OUTLOOK
Gene therapy

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Encryption keys for communication between multiple users PAGE 225

QUANTUM NETWORKING

BIOETHICS

HUMAN EMBRYO MODELS

Time for an international discussion to steer research

PAGE 183

ORGANIC CHEMISTRY

REFLECTION REMOVAL

Chiral allenes deracemized with visible light

PAGES 197 & 240

IMMUNOTHERAPY

ADRENALINE RUSH

How certain hormones help drive cytokine storms

PAGES 194 & 273

NATURE.COM
13 December 2018
Vol. 564, No. 7735

THIS WEEK

EDITORIALS

MEMORY Collective attention declines according to maths **p.162**



WORLD VIEW We must mentor students in their own languages **p.163**

MERCURY RISING Melting Arctic permafrost will release toxic metal **p.164**

Wanted: a fair carbon tax

Unrest in France at a rise in fuel prices highlights how the necessary transition to a clean economy must be carefully managed.

The principles of corporate change management stress that, although transitions must start at the top, the real change needs to happen at the bottom. That's a lesson that French President Emmanuel Macron perhaps wishes he'd remembered as protestors rioted in the Paris streets over the past few weeks against a planned new green tax that would have made fuel more expensive.

The movement has some support from economists, who tend to view the blanket introduction of such green taxes as regressive: the poorer people are, the greater the proportion of their income they spend on basics such as fuel, and so the heavier they find the burden when those goods are taxed. Hence the French 'Yellow Vest' protestors have complained — with some justification — that the new fuel tax places an unfair demand on those who can least afford it.

Events in France highlight the need for the 'just transition' that environmentalists and researchers have been pushing for a long time: smart climate policies must be fair, addressing both opportunities and inequalities.

In the long term, the benefits for humanity of a societal shift away from fossil fuels and towards cleaner sources of energy will far outweigh the costs. But the transition could have severe implications for some sectors, regions and countries. Poorly managed, it could result in loss of income, opportunity and future prospects for some workers and communities. So — and this is a question being discussed at the United Nations climate talks in Poland this week — how can it be managed well?

"One way to make a carbon tax more palatable to the taxpayers is to give them the money back."

Investment in renewable energy is making great strides and the cost of wind and solar energy is falling fast. But it is inevitable that the cost of fossil fuels in many applications will have to rise to force the pace of the transition to a cleaner economy. The surest way to do this is through some kind of carbon tax. (Global politics has turned firmly away from the other major route, a cap-and-trade system.) And one way to make a carbon tax more palatable to the taxpayers is to give them the money back.

That's essentially what Canada plans to do. Starting next year, Prime Minister Justin Trudeau's government will introduce a national 'fee and dividend' scheme that will place a levy on the carbon emissions of fuels and other products, but then refund the money to individuals and companies through tax rebates.

Most residents and businesses in Ontario, Saskatchewan, Manitoba and New Brunswick — the four provinces subject to the federal tax (other states have introduced their own versions) — will receive refunds that, the government says, will be greater than the carbon tax paid by the average family. According to the government's estimates, some 70% of people will get back more in dividends than they pay in new tax. Only those that use a lot of fuel will end up out of pocket. It's a bold move and one that will help to determine whether Trudeau remains in office after the general election scheduled for October.

The introduction of the French tax has now been suspended for six months, to give officials more time to ponder their response. Governments and policymakers elsewhere will be watching with interest. So will environmentalists and economists. If the question for the twentieth century was about the role of people in causing climate change, the conundrum now lies in finding a politically acceptable way to persuade or compel people to take the required action to reduce emissions. ■

Culture change

Improvements to a conference accused of sexism are long overdue.

The challenges and promise of artificial intelligence drew hundreds of scientists to a conference in Montreal, Canada, last week. But it was human behaviour that was the focus of much of the attention. The event and the board of trustees that oversees the conference have been in the spotlight in recent months over claims that previous gatherings had fostered a hostile environment for women.

Exhibit A is the acronym that the event commonly went by: NIPS. Although its defenders could say it merely reflected the full title of the organization — Neural Information Processing Systems — the board agreed to a last-minute change. So, this year, machine-learning researchers, software engineers and programmers arrived in Canada for the 'first' NeurIPS conference.

It's a small change, but a necessary, overdue and symbolic one. In a previous year, researchers attending a workshop for women in machine learning experienced boorish and offensive behaviour by some men who arrived wearing T-shirts emblazoned with a joke about nipples. And earlier this year, a survey of past attendees found that many respondents had experienced harassment, bullying and a lack of respect.

It is wrong that people ever experienced this behaviour, and it is sad that it has taken this long to respond, but the board deserves at least some credit for its response to the concerns raised by those in the community it represents, and for taking explicit steps to challenge and change the culture of the event.

The diversity and inclusion co-chairs of this year's organizing committee, for example, sent a strong message about the expected conduct of attendees when they discussed at the conference's opening remarks the measures in place at the event to make it more inclusive. The first invited talk also covered the necessity of diversity in technology.

It is difficult to know whether these and other actions had a measurable effect. But women who have attended in the past reported a welcome shift in the atmosphere of this year's event, and many applauded the board and organizers for their efforts to combat bad behaviour and

© 2018 Springer Nature Limited. All rights reserved.

encourage inclusiveness. These include an updated code of conduct that forbids event sponsors from using sexualized clothing or costumes, a town-hall meeting to discuss the issues, on-site childcare and stickers that help inclusion by flagging up first-time attendees and highlighting the pronoun that people prefer to be referred to by. In addition, more specific meetings for under-represented groups ran alongside the conference than in previous years.

They are small steps down a long road. Most participants easily adopted the name NeurIPS, with only a few accidentally slipping up and mentioning NIPS. There were no offensive T-shirts at conference events and no supply of commemorative coffee mugs — at least when the conference opened — as the name change came too late to get them printed.

Too often, the burden of work involved in increasing diversity and inclusivity falls on those from under-represented groups. NeurIPS is no exception.

One of the organizers of the Black in AI workshop at NeurIPS, Timnit Gebru, a computer-vision researcher at Stanford University, California, spoke for many of the session organizers when she told the diversity town-hall meeting that coordinating the event had reduced the time available for her research. The diversity and inclusion co-chairs, Katherine Heller and Hal Daumé, who have had to walk the fine line between a vocal research community pushing for change and a conference board that has been slow to realize the significance of its actions, also say they have seen considerable disruption to their research. Only Heller, who is at Duke University in Durham, North Carolina, has so far committed to returning to the post next year. These examples underscore the fact that increasing diversity is a job for everyone and it is not sustainable or fair to rely on a small number of volunteers to do this important work.

The challenge should not be underestimated. The organizers of another major AI conference, the International Conference on Learning Representation, announced last month that they would hold their 2020 event in Addis Ababa in a bid to widen the pool of talent that can attend. But this well-meaning initiative is not without problems. It is illegal to be gay in Ethiopia, raising questions over whether the Queer in AI workshop can take place there. The organizers are hoping to get express permission for the event from the Ethiopian government.

"Increasing diversity is a job for everyone, it is not fair to rely on a small number of volunteers."

The Canadian government has come under fire for denying entry or being slow to approve visas for many researchers invited to attend NeurIPS from overseas. More than half of the 200 people who sought visas to attend the Black in AI workshop did not receive them in time, including several who dedicated huge amounts of time to organizing the event.

There have been many high-profile criticisms of AI algorithms that mimic, and so perpetuate, the biases of wider society. That the board of an AI conference — which in October called in a diversity and inclusion consultant to assist it — has taken a stand against such discrimination within its own ranks is a necessary and overdue step. Ensuring changes are deep and lasting will take much more time and effort. Meanwhile, many more institutions and organizations need to follow. ■

How we forget

From pop music to tennis stars, society loses interest according to a mathematical law.

In his enthralling 2009 collection of parables, *Sum: Forty Tales from the Afterlives*, the neuroscientist David Eagleman describes a world in which a person only truly dies when they are forgotten. After their bodies have crumbled and they leave Earth, all deceased must wait in a lobby and are allowed to pass on only after someone says their name for the last time. "The whole place looks like an infinite airport waiting area," Eagleman writes. "But the company is terrific."

Most people leave just as their loved ones arrive — for it was only the loved ones who were still remembering. But the truly famous have to hang around for centuries; some, keen to be off, are with an "aching heart waiting for statues to fall".

Eagleman's tale is an interpretation of what psychologists and social scientists call collective memory. Continued and shared attention to people and events is important because it can help to shape identity — how individuals see themselves as part of a group — and because the choice of what to commemorate, and so remember, influences the structures and priorities of society.

This week in *Nature Human Behaviour*, researchers report a surprising discovery about collective memory: the pattern of its decay follows a mathematical law (C. Candia *et al. Nature Hum. Behav.* http://doi.org/cxq2; 2018). The attention we pay to academic papers, films, pop songs and tennis players decays in two distinct stages. In theory, the findings could help those who compete for society's continued attention — from politicians and companies to environmental campaigners — to find ways to stay in the public eye, or at least in the public's head.

The study applies maths and a big-data approach to questions that have been studied at length in the social sciences. Using attention as a proxy for memory, the authors analysed online views of the Wikipedia profiles of around 1,700 sports stars, citations of almost 500,000 physics papers and 1.7 million patents, and online play counts of some 33,000 songs and 15,000 film trailers.

Researchers had previously thought that the decline in the popularity of such cultural objects followed a smooth, steep curve. But analysis of the new study data revealed that a better fit was a shape called a biexponential function, which has two phases. It shows that collective memory dropped quickly, but that the subsequent decline in attention slowed considerably, and went down a much gentler slope. Although the shape was the same for each feature studied, the actual length of each phase was different. Music showed the shortest and sharpest initial decline in attention (taking 6 years) and the online biographies of the sports stars the longest (20–30 years).

How come? The researchers propose an explanation. The first, steep decline phase is dominated by the process of communicative memory, which is the direct word-of-mouth transfer of information. And the second, more enduring phase relies more on cultural memory, which is sustained by the physical recording of that same information.

That requires, of course, that the information is recorded. As an accompanying News & Views article highlights, for events that are memorialized with few cultural artefacts, such as Hurricane Sandy striking New York in 2012, policymakers could look at how to extend the period for which communicative memory dominates (A. Coman *Nature Hum. Behav.* http://doi.org/cxst; 2018). For a short time, conversations about the damage it caused probably raised awareness of climate change as a serious threat. But as collective memory of the severity of the hurricane faded, so, too, did concern.

The model does not apply in all cases, of course. Everyone will have their own example of an enduring figure still waiting in Eagleman's purgatorial lobby for their name to become redundant. But it's a neat way to apply the promise of big data to a new field of study, and one that could have real-world applications. It's also another example of how what can seem to be random and individual events when studied at a large enough scale can reveal an underlying pattern. The researchers compare their biexponential function of collective-memory decay to the more poetic description of a two-phase system from Chilean writer Pablo Neruda: "Love is so short, forgetting is so long." Which, at the very least, should keep Neruda hanging around for a bit longer. ■

© 2018 Springer Nature Limited. All rights reserved.



CLARISSA RIOS ROJAS

Students need guidance in languages they speak

Mentorship programmes in languages besides English could unlock opportunities for young scientists in Latin America, says **Clarissa Rios Rojas**.

Have you ever wondered what might happen if we unleashed all the professional talent going untapped in Latin America? There is a way to find out: boost mentorship programmes.

The 2017 World Bank report *Higher Education in Latin America and the Caribbean* finds that of people aged 25–29 who have ever been enrolled in higher education in the region, only half have finished their degrees. It attributes dropout rates in part to a "lack of mentoring, tutoring, and counseling programs" (see go.nature.com/2fy8dmr). Although there are many excellent mentorship programmes in high-income countries, most are offered only to students already in those nations, and usually only in English.

I come from a middle-income family in Peru. As an undergraduate student, it seemed unthinkable for me and my classmates to attend international conferences, go on internships abroad, gain access to top scientific journals or even meet scientists conducting research using state-of-the-art technology. (In 2006, DNA microarray analysis was mainstream; as far as I know, no one in Peru was using it.) I didn't know how to go about any of these tasks, and money was scarce. In my biology class at a public, national university, maybe 8 out of 60 students were able to afford English classes. Some struggled even to pay the university fee, which, when I was there, amounted to 80 Peruvian soles (US$24) a semester.

I was more fortunate than most in Peru's public-university system. I had parental support, and the rare chance to learn English at secondary school and while at university. Knowing English gave me access to most scientific literature, and opened up many opportunities. During the last semester of my bachelor's degree, I received one of three scholarships given by the University of Turku in Finland for university students in Peru for exchange studies (the programme is no longer running). Of the 30 or so scholarships I have received to study or travel to events, conferences and workshops, only two applications could be completed in Spanish.

During the final year of my PhD in Australia, I realized how lucky I was. I wanted to help Latin American students who are hampered by a lack of money and contacts. So, in 2015, I founded Ekpa'palek, a non-profit whose name means 'teaching a child to take their first steps' in Shiwilu, an Indigenous Amazonian language. It offers free online mentorship in Spanish and Quechua — the primary Indigenous language of Peru — and we are in the process of adding other languages.

Interested students can go to our webpage (www.ekpapalek.com) to find a suitable mentor: someone who works in their field, speaks their language, works in a country of interest or offers a particular set of skills. Currently, we host profiles of about 40 mentors (nearly all originally from Latin America), with expertise in the physical, biological and social sciences, life coaching and English writing. Mentors and mentees connect, sometimes once, sometimes for ongoing guidance.

Today, Ekpa'palek's blog is read in 116 countries, and the videos on our YouTube channel have been viewed more than 68,000 times in total. Some of our mentees are now studying or doing internships at top universities in the United States, Canada, Mexico and Brazil, among others.

The 174 students and young professionals who have so far participated have asked for help on tasks such as improving their CV, finding opportunities abroad, choosing a speciality, becoming internationally competitive and — inevitably — how to improve their English. We have had students who are eager to learn about careers not taught in some Latin American countries (Peru and Guatemala, for example, do not seem to offer any masters degrees in astrophysics). Some have sought advice on which countries are most welcoming to gay people, or how to deal with rape trauma. (Latin America is particularly prone to violence against women and people from sexual and gender minorities.) Mentoring programmes should be designed to help students not only with professional development, but also with their self-confidence, emotional intelligence and personal issues.

We also know there are more students in the developing world than we can reach. There are programmes, including AuthorAID and Cientificos.pe, that help young people who are already engaged in a career as scientists in Latin America. There are fewer that help students to learn which opportunities are available, and give guidance on how to access them. I have been unable to find any university or government department in Latin America with a solid mentorship programme.

Many universities globally host excellent mentorship programmes; Latin American universities should, too. Governments could alter curricula at national universities to include more soft skills. Organizations such as the United Nations Educational, Scientific and Cultural Organization could emphasize mentoring programmes for professional development in the developing world.

Virtual mentorship programmes such as Ekpa'palek can help to attenuate the brain drain that many Latin American countries face, by bringing back knowledge and connections, if not people. So can Serendipity, another organization created by scientists, which offers virtual mentorship for science, technology, engineering and mathemathics students in Peru.

But these are not enough, especially when both rely entirely on crowdfunding and voluntary work. More programmes — both virtual and institutional — are needed to unlock the vast talent and potential of Latin American students. ■

VIRTUAL MENTORSHIPS CAN HELP TO ATTENUATE THE BRAIN DRAIN THAT MANY LATIN AMERICAN COUNTRIES FACE.

Clarissa Rios Rojas *is founder and director of Ekpa'palek, a doctoral fellow at the Geneva Centre for Security Policy in Switzerland and co-lead in science advice at the Global Young Academy.*
e-mail: c.rios@gcsp.ch

© 2018 Springer Nature Limited. All rights reserved.

SEVEN DAYS *The news in brief*

EVENTS

Forests in 3D

A NASA instrument that will map the world's temperate and tropical forests in three dimensions launched to the International Space Station on 5 December. The Global Ecosystem Dynamics Investigation (GEDI) will use an advanced laser to analyse the heights of trees, shrubs and other foliage to gather information about the amount of carbon contained in Earth's forests. Scientists hope that GEDI will give them a much better picture of the planet's carbon sources and sinks. The instrument was launched aboard a SpaceX rocket with other scientific equipment, as well as 40 mice for a study about ageing. The launch, originally scheduled for 4 December, was delayed by a day to replace mouse food that had gone mouldy.

POLICY

Emissions limits

The US Environmental Protection Agency (EPA) has announced plans to weaken the greenhouse-gas emissions standards for new, modified or reconstructed power plants. The proposed changes, released on 6 December, would replace regulations that effectively require any new or substantially modified coal-fired power plants to be equipped with the technology to capture and store carbon dioxide emissions. Opponents of the regulations, put in place under then-president Barack Obama, have argued that carbon-capture technology is too expensive and not commercially viable. The plan is the latest attempt by the administration of President Donald Trump to roll back climate policies implemented under Obama. The administration has already sought to scale back greenhouse-gas emissions standards for existing power plants and cars.



XINHUA VIA ZUMA

Chang'e-4 sets off to the Moon's far side

China's Chang'e-4 spacecraft, which is bound for the Moon's far side, successfully lifted off from the Xichang Satellite Launch Center in Sichuan province on 8 December. The craft, carrying a lander and a rover, aims to be the first to 'soft' land on the Moon's crater-filled far side. The rover will survey its surroundings, and the lander will carry out several experiments, including testing whether plants can grow on the Moon. "Everything appears to have worked flawlessly," says Robert Wimmer-Schweingruber, a physicist at the University of Kiel, Germany, who has a radiation-detection experiment on the lander. Although the landing date has not yet been officially announced, Chang'e-4 is expected to attempt to touch down on the Moon's surface early next month. The site will probably be inside a 186-kilometre-wide crater called Von Kármán.

Moral code for AI

Around 530 people have signed up to a set of ethical guidelines for the development of artificial-intelligence (AI) technologies. The University of Montreal and the Quebec Research Fund spent a year devising the list of principles, known as the Montreal Declaration, with the help of the general public, sociologists and policymakers. The ten principles include respecting privacy and ensuring that systems are democratic, equitable, responsible and designed to make sure that everyone in society benefits from AI. Most of those who have signed the declaration work in Canada or France.

Space agency

The Philippines is a step closer to creating a national space agency. On 4 December, lawmakers in the country's lower house of congress unanimously approved a bill to establish an institution to shape national space policy. The bill now needs to pass a vote in the upper house. If created, the agency will focus on space science and technology applications that could address national issues, such as disaster-risk reduction. The space agency will be the country's official representation in the international space community. Although the Philippines already has space-related research and development initiatives, these are distributed between several national and private agencies. In 2015, the country established the National SPACE Development Program, led by astrophysicist Rogel Mari Sese, to jump-start the growth of space-science research. Since 2016, another programme to develop microsatellites has launched two into space to help map and monitor the

© 2018 Springer Nature Limited. All rights reserved.

Philippines for hazards, and it has launched a cube satellite for communications.

SPACE

Voyager milestone

NASA's Voyager 2 spacecraft has crossed into interstellar space, joining its twin — Voyager 1 — which made the passage in 2012. Both probes now sail beyond the reach of the Sun's influence, as humanity's most distant emissaries. Voyager 1 is 21.6 billion kilometres from the Sun. Voyager 2 is 18 billion kilometres away, and crossed the boundary on 5 November, said Ed Stone, Voyager's project scientist based at the California Institute of Technology in Pasadena. He made the announcement on 10 December at a meeting of the American Geophysical Union in Washington DC.

POLITICS

Detained scholar

A group of 121 Nobel laureates has written an open letter to Iran's supreme leader calling for the release of Ahmadreza Djalali, a disaster-medicine scholar who has been sentenced to death in the nation. The letter was distributed to participants of the



KEVIN VAN DEN PANHUYZEN/ZUMA

Nobel prize ceremony in Stockholm on 10 December by the human-rights group Amnesty International. It says that Djalali's health is declining rapidly. The researcher, who worked at the Karolinska Institute in Stockholm, was arrested in April 2016 on spying charges during a visit to Iran. He was sentenced after a trial in Iran's revolutionary court. Djalali (**pictured,** protestors holding his photo) denies the charges.

RESEARCH

Earthquake risk

Three new maps reveal which parts of the globe are most at risk of earthquakes — and where most people are vulnerable to seismic disaster. The charts, released on 5 December, are the culmination of a years-long effort coordinated by the Global Earthquake Model (GEM), a non-profit organization in Pavia, Italy, that works with emergency-management officials, geological surveys and disaster-preparedness groups worldwide. The first map, of global seismic hazard, shows which parts of the globe are prone to earthquakes. The second, of global seismic risk, highlights areas where buildings are likely to be damaged by ground shaking, and the third, of global exposure, looks at the number of buildings around the world — emphasizing the danger in highly populated regions.

PEOPLE

UK science minister

The UK government appointed a new science minister, Chris Skidmore, on 5 December, after Sam Gyimah resigned from the post over Brexit negotiations. Skidmore takes over responsibility for the universities and science portfolio, which is split between the education and business departments; he is the United Kingdom's fifth science minister since 2010. Skidmore represents a constituency in southwestern England that hosts a national research centre on composite science, and he served as an adviser to former science minister David Willetts. Gyimah resigned over Prime Minister Theresa May's politically divisive Brexit divorce deal, which defines the terms of Britain's looming exit from the European Union. On 10 December, May delayed a crucial vote on the agreement in the face of major opposition to the plans among members of the UK Parliament. She said that she hopes to amend the deal to address their concerns.

Topology pioneer

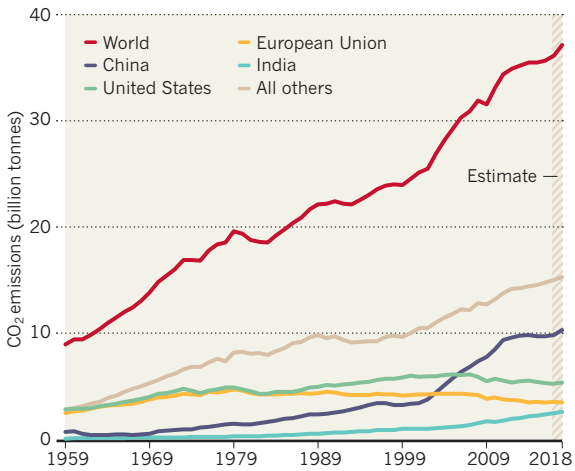
Theoretical physicist Shoucheng Zhang, a pioneer in the study of topological states of matter, died on 1 December, aged 55. Zhang, who was at Stanford University in California, was among the first physicists to predict that some materials known to be insulators should be able to conduct electricity on their outer surface. Such effects should arise because the quantum states of electrons can form shapes that are robust under perturbation, like knots on a string that can be pulled and twisted but not undone. These features are described by the mathematics of topology, and the materials are dubbed topological insulators. Zhang worked with others to confirm his predictions in the lab, and received several prizes for the work. Born in Shanghai, China, in 1963, he studied in Germany and the United States.

TREND WATCH

Global industrial emissions of carbon dioxide are likely to have risen by 2.7% in 2018 to reach an all-time high, an international consortium of scientists reports. This marks a second year of strong growth after a brief period of relatively stable emissions. The findings were released by the Global Carbon Project on 5 December at the 24th Conference of the Parties to the United Nations Framework Convention on Climate Change (COP24) in Katowice, Poland — and they underscore the challenge of reining in fossil-fuel consumption. The scientists also said that deployment of renewable energies such as wind and solar power is increasing rapidly around the world, but not fast enough to displace coal use in places such as India and China or a growing global demand for oil and natural gas. Industrial CO_2 emissions are likely to hit an all-time high of 37.1 billion tonnes this year. The total CO_2 emissions, including those from deforestation and other activities on land, will reach 41.5 billion tonnes — also the highest since records began. The biggest driver of emissions growth is a rise in fossil-fuel emissions in China, which accounts for more than 46% of the projected increase.

CO_2 EMISSIONS KEEP ON RISING

Industrial carbon dioxide emissions are projected to rise again globally this year, even as individual countries' emissions look very different.



SOURCE: GLOBAL CARBON PROJECT

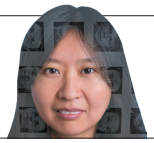
© 2018 Springer Nature Limited. All rights reserved.

NEWS IN FOCUS

HEALTH Mysterious polio-like disease probed with machine learning **p.170**

PUBLISHING China gets behind bold European open-access plan **p.171**

INFECTIOUS DISEASE Ebola cases slip under the radar **p.174**



NEUROSCIENCE Unlocking how the brain makes sense of faces **p.176**

PATRICK LANDMANN/SCIENCE PHOTO LIBRARY



The unique skeleton was embedded in rock deep inside a South African cave.

ARCHAEOLOGY

'Little Foot' fossil chiselled out of stone yields secrets

Mysterious ancient hominin retrieved after 20-year effort might be a distinct species.

BY COLIN BARRAS

After a tortuous 20-year-long excavation, an ancient skeleton is starting to reveal new information about early human evolution. The first of a raft of papers about 'Little Foot' suggests that the fossil is a female who showed some of the earliest signs of human-like, bipedal walking, around 3.67 million years ago. She might also belong to a distinct species unfamiliar to most researchers. "It's almost a miracle it's come out intact," says Robin Crompton, a musculoskeletal biologist at the University of Liverpool, UK, who collaborated with the research team that excavated the skeleton.

The nickname Little Foot, echoing the mythical 'Bigfoot', refers to the small foot bones that were among the first parts of the skeleton to be discovered. In 1994, Ronald Clarke, a palaeoanthropologist at the University of the Witwatersrand (Wits University) in Johannesburg, South Africa, was rifling through boxes of fossils at a field laboratory at the Sterkfontein caves, about 40 kilometres northwest of Johannesburg. He realized that a handful of small bones in the collection belonged to a species of *Australopithecus* — ape-like hominins in Africa between about 4 million and 2 million years ago, before the human genus *Homo* rose to dominance[1].

Clarke and his colleagues then found more bones embedded in a matrix of solid rock deep in the caves. They began carefully excavating Little Foot, piece by fragile piece, using hammers and chisels followed by precision tools. "The fossilized bone is actually softer than the matrix," says Crompton. "It's been an ▶

© 2018 Springer Nature Limited. All rights reserved.

absolute devil to get it out."

By late last year, Clarke's team had removed enough bones to reconstruct more than 90% of the skeleton — making it the most complete *Australopithecus* so far. On 29 November, they posted two papers on Little Foot to the bioRxiv preprint server — one on the age of the specimen[2], the other on the limbs and locomotion[3].

On 4–5 December, the team posted third and fourth papers, on the skull and the potential relationship of the specimen to a known hominin species[4], as well as on the arms and an injury Little Foot received during her life[5]. Further papers, on the hand, teeth and inner ear, are expected in the near future, says Crompton. Most will ultimately appear in a special edition of the *Journal of Human Evolution*.

A NEAR-COMPLETE PUZZLE

The bioRxiv papers crystallize ideas that emerged in earlier publications about the age of the fossil. They also cover new ground, suggesting that Little Foot was an adult female and stood about 130 centimetres tall — just 10 centimetres shorter than the average woman in some modern-human populations. "Little Foot was quite big," says Crompton. The paper covering limbs and locomotion[3] reveals that Little Foot's legs are longer than her arms, similar to modern humans, making her the oldest hominin for which we can be sure of that feature, says Crompton. This means that Little Foot was better adapted to walking upright on the ground than were many other australopiths.

Little Foot's skull, bones and teeth are so unusual that Clarke and his team have categorized her as the distinct species[4] *Australopithecus prometheus*, a name first suggested in 1948 on the basis of a skull fragment found roughly 250 kilometres north of Johannesburg[6] and that remains controversial. They also suggest that *A. prometheus* is an ancestor of a group of hominins called *Paranthropus*[4], which co-existed with early *Homo* species for about one million years.

But Lee Berger, an archaeologist also at Wits University, disagrees with the decision to resurrect *A. prometheus*. In a paper scheduled to be published in the *American Journal of Physical Anthropology*, he argues that the name *A. prometheus* was never properly defined. If Little Foot constitutes a distinct species, Berger thinks, a new name is needed.

He is also disappointed by the lack of solid information in the papers on the age and locomotion — he would have liked to have seen detailed measurements of the fossil bones. "There's no data — there are almost no measurements of the fossils," he says. Berger hopes to provide those data in his own publications — although he is still at an early stage of his analysis of Little Foot.

Crompton responds that the locomotion paper is an overview that attempts to reconstruct how Little Foot moved by drawing on the more-solid data in the team's other papers. Gabriele Macho, an anthropologist at the University of Oxford, UK, agrees that the locomotion paper is light on solid data, but says the team acknowledges the gap. She looks forward to seeing more-detailed papers soon. "The positive thing is this skeleton is tremendously important," she says. "No question about it." ■

1. Bruxelles, L., Clarke, R. J., Mairee, R., Ortega, R. & Stratford, D. *J. Hum. Evol.* **70**, 36–48 (2014).
2. Bruxelles, L. *et al.* Preprint at bioRxiv https://doi.org/10.1101/482711 (2018).
3. Crompton, R. H. *et al.* Preprint at bioRxiv https://doi.org/10.1101/481556 (2018).
4. Clarke, R. J. & Kuman, K. Preprint at bioRxiv https://doi.org/10.1101/483495 (2018).
5. Heile, A. J., Pickering, T. R., Heaton, J. L. & Clarke, R. J. Preprint at bioRxiv https://doi.org/10.1101/486076 (2018).
6. Dart, R. A. *Am. J. Phys. Anthropol.* **6**, 259–284 (1948).

GENETICS

Machine learning hunts for cause of paralysing illness

Scientists hope that probing the immune system will identify the cause of a polio-like disease.

BY SARA REARDON

Infectious-disease researchers hunting for the cause of a mysterious illness that is paralysing children are combining machine learning with a new gene-sequencing technique to pin down the culprit.

The disease, called acute flaccid myelitis (AFM), causes limb weakness and paralysis that resembles the symptoms of polio. The US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, has confirmed 134 cases of AFM in the United States so far this year. Many of those who develop the illness never recover.

Most of the evidence suggests that an enterovirus called EV-D68 is causing the illness[1], but researchers haven't been able to find the pathogen in the spinal fluid of children with the disease. Scientists are trying to identify the culprit by using a combination of host-response diagnostics — which look at how the immune system responds to pathogens — and machine-learning analysis. The approach could lead to better diagnostics and provide hints about new treatments.

Host-response diagnostic tests haven't been used in the clinic yet. But researchers are developing similar tests to help pinpoint other conditions that can be tricky to diagnose, including tuberculosis and bacterial meningitis.

"We've never really had a smoking gun."

This year's AFM outbreak started in October, and is the third in a series of outbreaks in the United States that have occurred every other year since 2014. Researchers have yet to find a definitive explanation for the pattern. It is also taking scientists an unusually long time to determine the cause of the illness, says William Weldon, a microbiologist at the CDC.

Blood samples taken from many of the people with AFM contain the virus. But many people who never developed AFM symptoms also have the virus in their blood.

"We've never really had a smoking gun," says Charles Chiu, an infectious-disease researcher at the University of California, San Francisco, who is leading the machine-learning project. He suspects that if EV-D68 causes AFM, it damages the spinal cord quickly and then drops to undetectable levels in the body.

Host-response diagnostics are useful when researchers don't know what they're looking for, says Purvesh Khatri, a computational systems immunologist at Stanford University in California. The composition of the immune system's defences differs depending on which pathogens are present in the body. So instead of looking for the agent itself, Khatri says, researchers could look at what the immune system is seeing.

Most attempts to identify mystery illnesses involve searching for a pathogen's DNA or RNA in areas of the body such as tissue or blood. But the host-response technique takes a blood sample and sequences all of the

© 2018 Springer Nature Limited. All rights reserved.

23,000 or so human genes present in the blood at any given time.

Chiu's group is analysing these genes — collectively known as the transcriptome — using machine learning. The scientists are searching for similarities between the transcriptomes of people with the illness, and differences between the transcriptomes of those with AFM and people with other, known infections, including those caused by enteroviruses. Once the team knows which genes are relevant to AFM cases, it can test for them directly.

"We're not relying on detecting the virus — we already know we can't detect the virus," says Chiu, who published some of the machine-learning methods in late November[2]. His group hasn't published any results yet because they're still preliminary. But the scientists' data suggest that the expressed genes that are common among people with AFM are those that researchers would expect to see in a person whose immune system is fighting a virus.

"I think it's definitely promising," says Weldon. He says that the CDC has been working with Chiu's group, and is talking with other teams that are pursuing similar experimental tests based on host-immune response.

Khatri stresses that researchers will need to train the machine-learning algorithm with data from diverse populations. Immune responses may vary depending on a person's ethnicity or country of origin, which can determine which pathogens people encounter, he says. Thorough training of the algorithm is especially important if researchers want to use similar host-response diagnostic techniques widely.

One group, led by infectious-disease researcher Christopher Woods at Duke University in Durham, North Carolina, has developed a transcriptomics test that can determine with 90% accuracy whether a bacterium, a virus or an autoimmune reaction is responsible for a person's symptoms[3].

The distinction is important for treatment, Woods says, and could prevent physicians from prescribing unnecessary antibiotics for viral or autoimmune diseases.

Khatri's group has developed a test that predicts whether a person will develop active tuberculosis. About 25% of the world's population harbours the bacterium that causes the illness, but only about 5–10% of these people develop symptoms[4]. The test from Khatri's group could allow researchers to categorize and prioritize people for treatment before the disease becomes severe.

Chiu hopes that the host-immune response approach could also help to explain why only some people infected with EV-D68 develop AFM. His group is also sequencing the genomes of children with the condition. The researchers hope that this information — combined with the transcriptome data — might provide hints about who could be susceptible to the illness before the next outbreak, which many researchers expect to occur in 2020. "These cases this year provide valuable data for us in evaluating how it might progress in the future if we see additional outbreaks," Chiu says. ■



ARMANDO L. SANCHEZ/CHICAGO TRIBUNE/TNS VIA GETTY

Most of those affected by the outbreak of acute flaccid myelitis are children.

1. Hixon, A. M. *et al. PLoS Pathog.* **13**, e1006199 (2017).
2. Langelier, C. *et al. Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1809700115 (2018).
3. Tsalik, E. L. *et al. Sci. Transl. Med.* **8**, 322ra11 (2016).
4. Sweeney, T. E. *et al. Lancet Respir. Med.* **4**, 213–224 (2016).

PUBLISHING

China backs open-access plan

Officials pledge support for 'Plan S', which aims to make papers immediately free to read.

BY QUIRIN SCHIERMEIER

In a huge boost to the open-access movement, librarians and funders in China have said that they intend to make the results of publicly funded research free to read immediately on publication.

The move, announced at an open-access meeting last week in Berlin, includes a pledge of support for Plan S, a bold initiative launched in September by a group of European funders to ensure that, by 2020, their scientists make papers immediately open.

It is not yet clear when Chinese organizations will begin implementing new policies, or whether they will adopt all of Plan S's details, but Robert-Jan Smits, the chief architect of Plan S, says the stance is a ringing endorsement for his initiative. "This is a crucial step forward for the global open-access movement," he says. "We knew China was reflecting to join us — but that it would join us so soon and unambiguously is an enormous surprise."

In three position papers, China's National Science Library (NSL), its National Science and Technology Library (NSTL) and the Natural Science Foundation of China (NSFC), a major research funder, all said that they support the efforts of Plan S "to transform, as soon as possible, research papers from publicly funded projects into immediate open access after publication, and we support a wide range of flexible and inclusive measures to achieve this goal". They add: "We demand that publishers should not increase their subscription prices on the grounds of the transformation from subscription journals to open access publishing."

The government will now urge Chinese funders, research organizations and academic libraries to make the outcomes of publicly funded research free to read and share as soon as possible, says Xiaolin Zhang, chair of the ▶

© 2018 Springer Nature Limited. All rights reserved.

Strategic Planning Committee of the NSTL at the Ministry of Science and Technology in Beijing. He told the meeting that the NSFC, NSTL and NSL will support the government's request to make research papers open immediately after publishing, and that implementation policies should follow soon. He expects funders to push all researchers in China to follow suit.

Zhang told the Open Access 2020 conference, convened by Germany's Max Planck Society, that any idea that open access has little traction in China is misleading. Since 2014, funders and research institutions in China have encouraged — and funded — scientists to publish papers in open-access formats, and to archive manuscripts openly online. But, he added, much of China's scientific output is locked behind paywalls. "NSFC funds about 70% of Chinese research articles published in international journals, but China has to buy these back with full and high prices," he says. "This is simply wrong — economically and politically."

He called on publishers at the meeting to start negotiating transformative deals with Chinese library consortia without delay. Such 'read and publish' agreements, which have been struck by a number of European national library consortia, and which the University of California system is also hoping to negotiate, cover the subscription costs of paywalled journals, but also allow corresponding authors at eligible institutions to publish their work openly in those journals.

CLEAR SIGNAL

China's commitment to ending subscription publishing took publishers at the meeting by surprise. "This is the first clear signal I received from China on this matter," said Daniel Ropers, chief executive of Springer Nature. "We were under the impression that open access isn't quite as urgent an issue in China as it is in Europe and the United States. If it is indeed, we are more than happy to engage."

Springer Nature, he says, already offers a broad range of open-access journals and would consider developing the portfolio further in all disciplines of science. But he says a viable solution is still needed for highly selective subscription journals, including *Nature*, to satisfy Plan S. (*Nature*'s news team is editorially independent of its publisher, Springer Nature.)

As it stands, the plan would bar scientists funded by participating agencies from publishing their work behind a paywall after 2020, unless they can also archive the accepted manuscript immediately online with a liberal publishing licence (which few subscription journals permit). Many subscription journals do offer an open-access option, but Plan S will fund publication by that 'hybrid' route in only some cases, and will review this policy in 2023. ■



The European Union's Galileo network (artist's impression) is a global satellite-navigation system.

BREXIT

UK satnav plan faces high hurdles

Britain says it has abandoned plans to rejoin the Galileo system for defence — but its alternative is problematic.

BY DECLAN BUTLER

The row over Britain's attempt to stay fully involved in the European Union's global satellite-navigation (satnav) system, Galileo, after it departs the bloc, is back in the headlines after science minister Sam Gyimah cited it in his resignation statement last month. Gyimah's resignation came after the country's Prime Minister Theresa May had said that the UK government would end talks with the EU on Galileo, and would instead consider building its own global satnav system for use after Brexit.

That idea was first floated by the government in May, but many experts have dismissed it as expensive, unnecessary and even unfeasible — the lack of available space on the radio spectrum to run such a system could be a show-stopper.

Nature digs into the dispute.

What did the science minister say about Galileo?

Gyimah said that the EU's superior hand in negotiations over the programme convinced him that Britain would fare badly in future Brexit negotiations on other issues, including research.

What is Galileo, and why is it so important?

Galileo is one of four global satnav systems, which provide myriad civilian, scientific and defence services. The others are the US Global Positioning System (GPS), Russia's Global Navigation Satellite System (GLONASS) and China's BeiDou, which will be fully operational in 2020. The EU started the Galileo programme in 1999 to break its dependence on the GPS and GLONASS.

The Galileo constellation — comprising 26 satellites — was completed this July; a near-complete constellation began beaming down

© 2018 Springer Nature Limited. All rights reserved.

DPA PICTURE ALLIANCE/ALAMY

signals free of charge to smartphones and other receivers in December 2016.

Researchers also combine signals and use them in an array of scientific applications, including the monitoring of movements in Earth's crust and for the study of the atmosphere.

The Galileo programme is building another 12 satellites as in-orbit spares and to replace older machinery. It is also starting to build a next-generation system. The EU opened the first of the tenders for building these craft in June. Total costs for the Galileo programme are estimated at around €13 billion (US$15 billion) to €15 billion to the end of 2020.

How would Brexit change the United Kingdom's participation in Galileo, and why is the UK government unhappy?

Brexit would have no effect on the availability of Galileo signals to scientists and other UK citizens — the service is freely available to anyone on the planet.

But a UK-based company, Surrey Satellite Technology in Guildford, a subsidiary of the aerospace giant Airbus, built all the satellites made so far (although many components, such as the satellites' atomic clocks, are sourced from suppliers in Europe).

However, the EU has already effectively excluded UK companies from bidding for the lucrative tender for the next-generation satellites. The British government has complained that this treatment is unfair, given its contributions so far.

After Britain leaves the bloc on 29 March 2019, it will also automatically stop being involved in the defence-related aspects of the Galileo programme — something the government was pushing to stay a part of.

What are Galileo's defence applications?

The system's secure service, scheduled to be fully operational by around 2026, will be restricted to government-authorized users, including the military and essential services such as energy supplies and telecoms. The signals are encrypted to stop interference or malicious jamming.

"Spending £3 billion to £5 billion on a UK system would be grotesquely wasteful."

The United Kingdom has been closely involved in the secure system's development. It had argued that this close participation, and its significant role in EU defence matters, mean it should be given special treatment that would allow it a full role in the inner workings of Galileo's defence aspects. But EU rules do not allow a non-member state to be involved in the development of such security aspects.

The United Kingdom said that this is unacceptable, leading May to say on 1 December that the government would abandon plans to use Galileo for defence and critical national infrastructure. She also confirmed that the United Kingdom was looking at options for building its own global system.

Is that proposal credible?

It might be technically feasible, say experts — Britain has the science and engineering skills to build such a system — but it probably isn't affordable. Widely cited estimates put the construction cost at somewhere between £3 billion (US$4 billion) and £5 billion. That doesn't include the running costs, which amount to about €800 million a year for Galileo. For comparison, the UK space agency's budget this year is £402 million, and Britain's defence research budget will be about £1.9 billion next year.

"Spending £3 billion to £5 billion on a UK system would be grotesquely wasteful," says Robert Massey, deputy executive director of the UK Royal Astronomical Society in London.

And even if Britain were to build its own system, there could be a crucial technical limitation: the lack of available space on the radio spectrum.

What's the issue with the radio spectrum?

The four existing global satnav systems already take up the part of the spectrum allocated for satellite navigation by the International Telecommunication Union (ITU), says Alexandre Vallet, head of the ITU's Space Services Department in Geneva, Switzerland. Squeezing in a new global system might require novel radio-signal designs that don't interfere with other systems, says Vallet. And these would need to be endorsed by international agreements — so it would be a challenge, he says. ■

CANCER RESEARCH

'Super' DNA targeted by drugs

DNA segments that amplify gene activity might represent a new form of gene regulation.

BY HEIDI LEDFORD

Experimental cancer treatments that harness souped-up segments of DNA called super-enhancers to activate genes are working their way to the clinic for the first time. But scientists are still debating how these elements work — and whether they represent a fundamentally new way of regulating genes.

Preliminary data suggest that screening for a particular super-enhancer can identify people with acute myeloid leukaemia who might benefit from a drug called tamibarotene. The data were presented by the drug's maker, Syros Pharmaceuticals, on 2 December at a meeting of the American Society of Hematology in San Diego, California. And on 15 November, the company debuted data from another preliminary trial, in which people with solid tumours were given a drug that targets a protein called CDK7. Laboratory tests have shown that inhibiting this protein can reduce the activity of a super-enhancer that has been linked to some cancers (E. Chipumuro *et al. Cell* **159**, 1126–1139; 2014).

The trials are the first attempts to target super-enhancers to treat human disease. But it is still unclear whether these DNA segments are truly stronger versions of better-known gene-regulating sequences called enhancers. "The word is still out," says Lothar Hennighausen, a geneticist at the US National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland. "I'm inclined to think that they are not."

Researchers have long known that enhancers are important for regulating when and how strongly genes are expressed. But in 2013, a group found that some enhancers, called super-enhancers, cluster together near genes that help to determine a cell's unique identity — whether it becomes a mammary or a muscle cell, for instance (D. Hnisz *et al. Cell* **155**, 934–947; 2013).

Super-enhancers seem to be particularly important in embryonic stem cells, and they are sometimes hijacked by cancer cells to drive the aberrant gene activity that fuels tumour growth.

And super-enhancers also attract unusually large numbers of the proteins required to activate the genes they control. These clusters of enhancers and proteins might allow cells to tightly regulate important genes, ensuring that they will be turned on exactly when needed and in precisely the right amount, says Christopher Vakoc, who studies gene expression at Cold Spring Harbor Laboratory in New York and has advised Syros.

"It's all about precision," says Vakoc. "When the cell goes to that much effort to control a gene, it's because the product of that gene is pivotal in biology."

Although mammalian cells have ▶

© 2018 Springer Nature Limited. All rights reserved.

thousands of enhancers, they typically have only a few hundred super-enhancers. As a result, researchers now use super-enhancers as a signpost for important genes, says Hennighausen. Understanding how they work could shed light on how cells adopt their identities. But researchers don't know whether the enhancers in a cluster act independently, or whether they work together in a new form of gene regulation.

That question arose right from the start, says Richard Young, a biologist at the Whitehead Institute for Biomedical Research and a co-founder of Syros — both in Cambridge, Massachusetts. "There were investigators who questioned whether or not they should have the term 'super', because it implied some function that typical enhancers didn't have," he says. "To be frank, at the time we didn't know if they had some special function."

Since then, researchers have scrutinized a few super-enhancers, studying the function of each enhancer in the cluster. But the results are inconclusive: some enhancers show signs of working together, whereas others seem to work independently. "It's a very intense debate," says Denes Hnisz, a molecular biologist at the Max Planck Institute for Molecular Genetics in Berlin.

Hnisz notes that the discrepancy might arise in part from the algorithms used to identify enhancers in genomic data: the algorithms could be mislabelling some sequences as super-enhancers. And different labs use different assays to test for super-enhancer activity, he adds, which could introduce another source of conflict.

Resolving the debate might have to wait until more scientists have studied more super-enhancers, says Douglas Higgs, a haematologist at the University of Oxford, UK. "At the current time, it is hard to be sure if they represent a new type of fundamental regulatory element."

For Syros, the debate is largely academic, says Nancy Simonian, the company's president and chief executive. "From our point of view, it doesn't really matter," she says. "We're just saying it's a marker for a hotspot that we know is associated with genes that are really important for controlling the cell."

The next few years could bring some answers. Studies of enhancers fell out of favour in the early 2000s, says Hennighausen. But technological advances are bringing them back into fashion. The ability to use relatively simple gene-editing tools, such as CRISPR–Cas9, to alter enhancer sequences has made it easier to study their function, he notes. An experiment that once took two years can now be done in a few months for much less money.

"The questions were always there, but the technology was needed to answer them," he says. "The whole field is emerging right now." ■

EPIDEMIOLOGY

Scientists seek hidden sources of Ebola

Two-thirds of new infections in the Democratic Republic of the Congo cannot be linked to known cases.



GORAN TOMASEVIC/REUTERS

An Ebola health worker carries a child at a hospital in the Democratic Republic of the Congo.

BY AMY MAXMEN

As the epicentre of the Ebola outbreak in the Democratic Republic of the Congo (DRC) shifts into the war-weary city of Butembo, public-health workers are trying to stamp out new infections from an unforeseen source: unregulated health centres.

Decades of political instability in the north-eastern DRC, the site of the epidemic, have fostered an increase in informal clinics that offer traditional and modern medicine. These centres treat people for malaria and other common illnesses, filling the vacuum left by the lack of a functional health system. But they are not designed to prevent the spread of a virus as dangerous as Ebola — which has put their patrons at high risk, according to the World Health Organization (WHO). Health officials have begun trying to lessen the centres' load by pre-emptively giving out malaria medication.

The push highlights a central challenge to ending the epidemic, which is now the second-largest on record: although experimental drugs and a vaccine have helped to limit Ebola's reach, two-thirds of new infections cannot be linked to existing cases. That has left epidemiologists racing to identify overlooked routes of infection. And conventional prevention measures have been thwarted by conditions in the north-eastern DRC, where decades of severe conflict have left millions of people dead, and millions more traumatized and homeless.

"I have lived through many outbreaks, but this is the worst one," says Jean-Jacques Muyembe-Tamfum, director-general of the National Institute for Biomedical Research in Kinshasa.

Already, 494 people have been infected with the virus, and 283 of those have died, the WHO said on 10 December. "This is as tough and complex as it gets," says Peter Salama, head of the WHO's health-emergency programme in Geneva, Switzerland.

In the city of Beni in North Kivu province, health workers have been struggling to identify and monitor people who might have been touched by someone in the throes of an

© 2018 Springer Nature Limited. All rights reserved.

infection. This work is crucial to containing the epidemic, but many people with Ebola will not name those they might have infected, says Annick Antierens, a strategic adviser at the humanitarian group Médecins Sans Frontières (also known as Doctors Without Borders).

"People are afraid and mistrust the system, so the contact lists are not very good," Antierens says. And some people with Ebola die without seeking medical help, or receive care only after they are too ill to communicate.

As a result, 66% of Ebola cases are springing up outside known chains of transmission. "When you look at a proportion like that, this outbreak should really be out of control," says Salama. But, on average, each person with Ebola in Beni is infecting just one other person, a success that Salama attributes to extensive vaccination. More than 40,000 people have received the experimental rVSV-ZEBOV Ebola vaccine since 8 August.

Analyses of patient data exposed the unregulated health centres as one unexpected source of cases, the WHO says. During some weeks in September and October, more than half of new Ebola infections were in children under the age of 16. Epidemiologists deduced that parents were taking their feverish children to the health centres for treatment, unwittingly exposing them to Ebola.

For years, the informal clinics have served communities that lack reliable, regulated healthcare facilities. Caretakers rarely have sterile equipment or even enough beds for patients, who must share, says Ibrahima Socé-Fall, director of the WHO's emergency operations for Africa, in Brazzaville in the neighbouring Republic of Congo. "Someone who has malaria will go to these facilities and drink from the same cup as the patient before them, or get an injection with the same needle," he says. "But you cannot tell people not to go there, since there is no alternative."

"I have lived through many outbreaks, but this is the worst one."

Instead, the WHO, the DRC government and aid organizations are attempting to support the clinics. They offer anyone who treats people — including traditional healers — the Ebola vaccine and training on basic measures for preventing the spread of infection, such as isolating people who are vomiting blood or showing other signs of Ebola. The task is enormous. There are an estimated 300 unregistered clinics in Beni alone, and outreach takes time, Antierens says: "We aren't just dropping off soap."

Ebola responders are also distributing anti-malaria drugs broadly across Beni. The DRC saw rising rates of malaria in 2017, with 435,000 deaths from the disease.

But greater challenges lie ahead as the centre of the outbreak moves from Beni to Butembo, an even more dangerous part of North Kivu. After years of arson, rape, murder and hunger, some communities in the area question the intentions of outsiders who are there to fight Ebola, Salama says. And some people distrust treatment centres because roughly half of their patients die in the first couple of days — in part, Antierens says, because many arrive too late for help.

Violence has also hindered response efforts directly. The WHO evacuated some staff members after a shell hit the hotel where they were staying in Beni on 16 November. Since then, Ebola responders who take temperatures and search for patients' contacts have been threatened. For their own safety, doctors must leave Ebola-treatment centres at sunset, says Muyembe-Tamfum. "We lose patients because they die in the night," he adds.

Political rallies organized in advance of the DRC presidential election, set for 23 December, could further complicate the Ebola response, by sparking violence and drawing people with the virus to mingle with those who do not carry it.

Salama predicts that the outbreak will continue for at least another six months. "I think we can stop this as long as security holds," he says, "but that's the big 'if'." ■

npj | Precision Oncology

Call for Papers

Exploring all areas of precision oncology



npj Precision Oncology is an online-only, open access journal, dedicated to publishing high-quality research covering all aspects of precision oncology from basic science to translational applications to clinical medicine.

Publishing with *npj Precision Oncology* offers authors a number of benefits, including:

- **High visibility** and **wide dissemination**
- **Global reach** and **discoverability** via nature.com
- **Compliance** with open access funding mandates
- **Comprehensive** and **rigorous peer review** by experts in your field

EDITORS-IN-CHIEF

Dr. Zigang Dong, MD, PhD, DrPH
Executive Director, The Hormel Institute, University of Minnesota, Austin, MN, USA; US China (Henan) Hormel Cancer Institute, Zhengzhou, China

Dr. Ann Bode, PhD
Associate Director and Professor, The Hormel Institute, University of Minnesota, Austin, MN, USA

Published in partnership with
THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

With support from
中美(河南)荷美尔肿瘤研究院
China-US(Henan)Hormel Cancer Institute

Part of the Nature Partner Journals series
npj nature partner journals

nature.com/npjprecisiononcology

natureresearch

A60191

© 2018 Springer Nature Limited. All rights reserved.

The face detective

Doris Tsao unlocked the brain's code for recognizing faces. Now she wants to work out how we see everything else.

BY ALISON ABBOTT

Doris Tsao launched her career deciphering faces — but for a few weeks in September, she struggled to control the expression on her own. Tsao had just won a MacArthur Foundation 'genius' award, an honour that comes with more than half a million dollars to use however the recipient wants. But she was sworn to secrecy — even when the foundation sent a film crew to her laboratory at the California Institute of Technology (Caltech) in Pasadena. Thrilled and embarrassed at the same time, she had to invent an explanation, all while keeping her face in check.

It was her work on faces that won Tsao awards and acclaim. Last year, she cracked the code that the brain uses to recognize faces from a multitude of minuscule differences in shapes, distances between features, tones and textures. The simplicity of the coding surprised and impressed the neuroscience community.

"Her work has been transformative," says Tom Mrsic-Flogel, director of the Sainsbury Wellcome Centre for Neural Circuits and Behaviour at University College London.

But Tsao doesn't want to be remembered just as the scientist who discovered the face code. It is a means to an end, she says, a good tool for approaching the question that really interests her: how does the brain build up a complete, coherent model of the world by filling in gaps in perception? "This idea has an elegant mathematical formulation," she says, but it has been notoriously hard to put to the test. Tsao now has an idea of how to begin.

Her ambitions for unlocking some of the most recalcitrant mysteries of the mind are no surprise to neuroscientist Margaret Livingstone, who advised Tsao throughout her PhD at Harvard Medical School in Boston, Massachusetts. "Doris never got sidetracked," she recalls. "She was quiet and focused, and always went for the big questions."

Tsao grew up in a household filled with science. Her mother worked as a computer programmer and her father was a machine-vision researcher. They emigrated to the United States from Changzhou, China, when Tsao was just four, "for a better life with more opportunities", she says.

"My father is probably the key reason why I study vision, though I try to deny it," Tsao says. Back when she was in high school, they discussed mathematical theories for how the brain might process aspects of vision. She found them "incredibly beautiful", she says. "He helped plant in my head the idea that vision requires a profound explanation."

She graduated in mathematics and biology at Caltech before joining Livingstone's team in 1996, where she initially studied the way the brain perceives depth of vision.

THE FACE CODE

Livingstone's lab works with macaques, which have a similar visual system and brain organization to those of humans. The view of the world through any primate's eyes is funnelled from the retina into the visual cortex, the various layers of which do the initial processing of incoming information. At first, it's little more than pixels of dark or bright colours, but within 100 milliseconds the information zaps through a network of brain areas for further processing to generate a consciously recognized, 3D landscape with numerous objects moving around in it.

During most of her PhD, Tsao was focused on the outermost layers of the visual cortex, where information from the retina first arrives. She learnt how to insert tiny electrodes — sensitive enough to record the firing of single brain cells — into this area of the monkeys' brains. But to help her probe deeper into the visual cortex, she decided to add brain imaging to her repertoire. The broader maps of brain activation provided by functional magnetic resonance imaging (fMRI) could help guide the more-precise single-cell recording techniques. Few labs at the time were imaging the brains of animals, but Wim Vanduffel, a pioneer of monkey fMRI at the Catholic University of Leuven, Belgium, helped Tsao to establish the infrastructure needed to do the work in Boston.

While learning about the technique, she became aware of a surprising fMRI discovery made by neuroscientist Nancy Kanwisher from the nearby Massachusetts Institute of Technology. Kanwisher had identified a small area of the brain in humans that lights up whenever a person is shown a picture of a face, but not when they are shown pictures of other objects such as a house or a spoon.

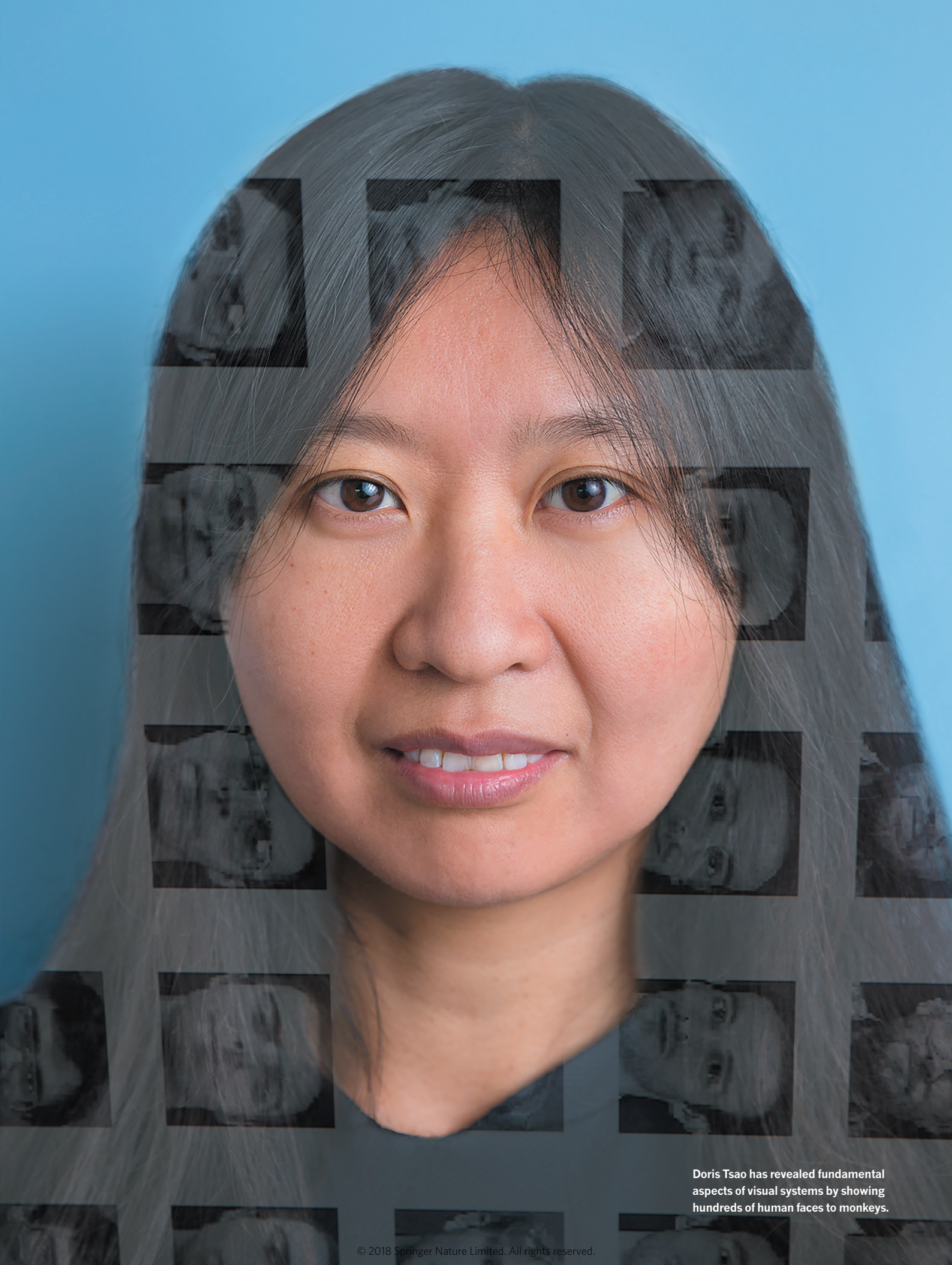
Tsao reasoned that if the same face-recognition system existed in monkeys, she could use her sensitive electrodes to probe the neurons involved and work out how they function.

She teamed up with Winrich Freiwald, who was then a postdoc in Kanwisher's lab, and began a series of experiments combining fMRI with single-cell recording techniques to probe the inferior temporal (IT) cortex, the brain region that Kanwisher had identified. Over the next eight years or so, Freiwald, Tsao and their collaborators made a number of important discoveries[1–3]. Passing picture after picture in front of the macaques, they mapped out the individual cells that fired in response to a human or monkey face. This allowed them to identify six patches on each side of the brain, distributed along the IT cortex. If the researchers electrically stimulated any one of the patches, the others lit up. Seeing those face patches working together in a network for the first time "was a joyful moment", says Freiwald, who is now at Rockefeller University in New York City.

Freiwald and Tsao also discovered that the patches tended to be specialized. By showing monkeys a series of cartoon faces with various details such as hair, a nose or irises missing, they could determine which cells fire in response to specific facial features. A cell's rate of firing

PHOTO ILLUSTRATION BY SAM COMEN FOR *NATURE*

© 2018 Springer Nature Limited. All rights reserved.



Doris Tsao has revealed fundamental aspects of visual systems by showing hundreds of human faces to monkeys.

© 2018 Springer Nature Limited. All rights reserved.

DECODING THE FACE

To decode facial recognition in monkeys, Doris Tsao and Steven Le Chang recorded signals from neurons in brain areas called face patches, while showing the animals hundreds of pictures of human faces. Face patches are located in the inferior temporal (IT) cortex, which is involved in visual processing of objects.

Parameterized face images

SHAPE DIFFERENCES

Specific cells in a patch at the top of the IT cortex respond to a given set of shape features, such as the space between the eyes, the width of a mouth or the shape of the hairline. These cells tend to respond to faces that are in one orientation.



Differing shape

APPEARANCE DIFFERENCES

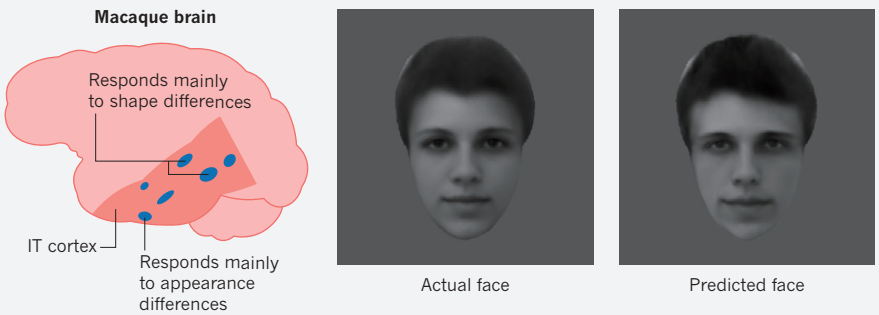
Cells in a patch found deeper in the IT cortex respond to appearance features, such as skin tone and texture. Some respond to the same faces regardless of orientation.



Differing appearance

PROJECTION TO PREDICTION

The degree of difference between facial features dictates the firing rate of neurons that respond. When researchers put together the neural activity about face shape and appearance from just 205 neurons, they could predict the features of a face the monkey was looking at. New work suggests that this type of patch organization, together with firing-rate responses, could underlie the mechanism by which other objects are recognized in the IT cortex.



PHOTOS: STEVEN LE CHANG AND DORIS TSAO

would ramp up according to how extreme the feature is, a property known as ramp-shaped tuning that turned out to be fundamental for face coding. A cell responding to the distance between two eyes, for example, might fire slowly in response to close-set eyes, but rapidly to ones set farther apart. When they showed monkeys real faces that were looking in different directions, the researchers discovered that cells in the patches closest to the visual cortex tended to fire in response to specific orientations of any face, whereas those in the deepest patch responded to a few individual faces, no matter what their orientations.

To investigate how the IT cortex might be encoding full faces from this information, Tsao realized that every face could be created by mixing the most important dimensions of 'faceness', such as how pointy a nose is, how eyes are set or complexion. She and her postdoc Steven Le Chang identified the 50 dimensions that varied most across faces — 25 for shape and 25 for appearance — and created a set of 2,000 face images in which the value of all 50 dimensions was known[4]. They flashed these images in front of the monkeys while measuring responses from 205 neurons in two face patches. The code started to reveal itself.

Cells in the more superficial patch tended to be tuned to shape dimensions, whereas many of those located deeper in the IT cortex responded to appearance dimensions. This made sense because the deeper cells might have to account for distorted shape dimensions when a head is turned. Tsao and Chang could predict how the neurons would fire on the basis of the dimensions of any face, and they could even reconstruct a face just from the firing patterns of these cells (see 'Decoding the face').

The research seemed to point to a mechanism by which individual cells in the cortex interpret increasingly complex visual information, until, at the deepest points, individual cells code for particular people.

That idea made intuitive sense. In 2005, Rodrigo Quian Quiroga, then a postdoc at Caltech, had identified what became known as Jennifer Aniston cells. By working with people who had had electrodes implanted in their brains to treat epileptic seizures, Quian Quiroga found signals from single neurons that responded to pictures of familiar or famous people. The cells also responded to any concept of that person. For example, one neuron fired in response to a photograph of the actor Jennifer Aniston, but also to her written name or even the title of a film she had starred in. These 'concept' cells resided in the hippocampus, which lies a little deeper in the brain than does the IT cortex[5].

Tsao met Quian Quiroga, now at the University of Leicester, UK, in 2015 at a small meeting in Ascona, Switzerland, where she was presenting her latest results. Over dinner, he asked her how she thought her face cells related to his concept cells. "They are probably their precursors," she told him. But she fretted about her answer throughout the night. One thing had always bothered her. The deep IT cortical cells that she had been working on often fired in response to several individual faces — those that didn't look like each other at all.

Unable to sleep that night, she thought through the mathematical analysis that she and Chang had been applying to their data. Then a moment of insight struck. She had gone over the maths that so neatly described the ramp-shaped tuning responses of cells a million times. But in the dark, silent hotel room, she realized that it was the same as a mathematical operation that describes a type of projection. Projection explains, for example, how the Sun might cast the same shadow for two different objects depending on how they are positioned. If the cells are simply projecting combined dimensions from a multidimensional 'face space', she says, "it would explain why lots of different faces could elicit the same response in a face cell". The IT cortex is not homing in on one particular person at all; that transformation must happen at a point even deeper in the brain.

A CATEGORICAL CHANGE

At breakfast, she told Quian Quiroga about her new hunch and found that he had been thinking the same thing. So, she made an unusual wager: she bet him a bottle of expensive wine that it would be wrong, "because if it were true, I would be happy without wine".

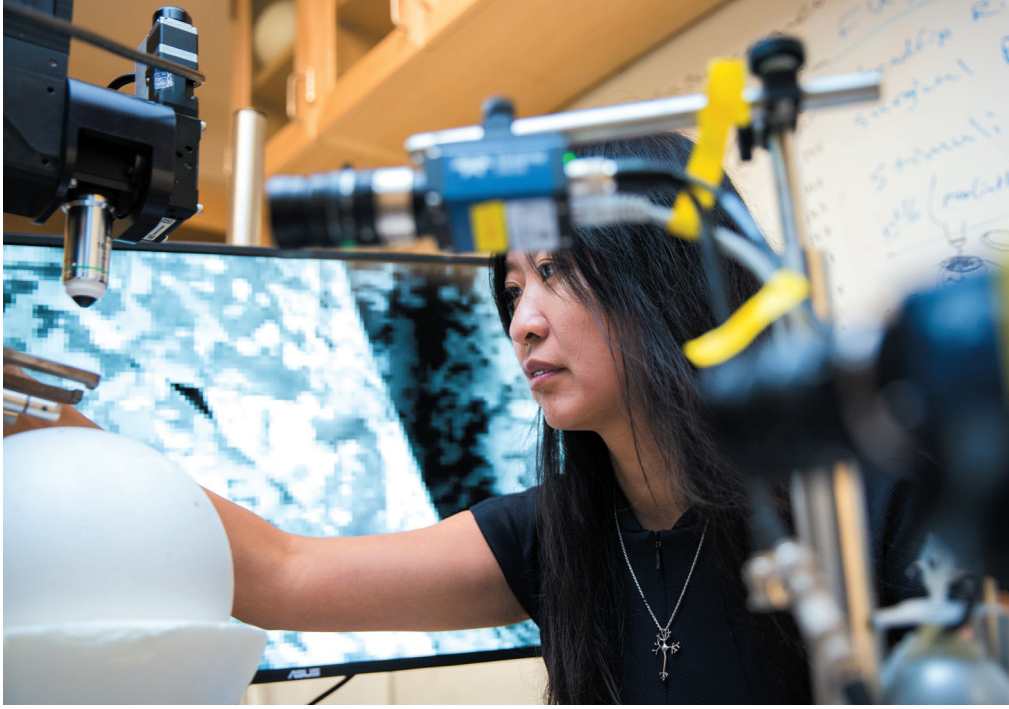
Rushing back to the lab, she and Chang embarked on additional experiments that lost her the bottle, but culminated in the publication[4] of the facial-recognition code in 2017.

The code was thrillingly — perhaps just a touch disappointingly — simple, says Tsao. That realization "was one of the happiest moments for me", she says.

There is a good chance that the same

© 2018 Springer Nature Limited. All rights reserved.

SAM COMEN FOR *NATURE*



Combining brain imaging and electrophysiology has helped Doris Tsao to peer deep into the primate brain.

simple code might apply over the whole of the IT cortex. Scientists have discovered other networks similar to the face-patch network that respond to other things, including bodies[6], scenes[7] and coloured objects[8]. But most of the IT cortex is uncharted territory. At a neuroscience meeting in Berlin this summer, Tsao presented some details of her current work. With her postdoc Pinglei Bao, she electrically stimulated cells in what she calls the no-man's land of the IT cortex, while scanning the monkey's brain. Two patches lit up, indicating another network — but this time she had no idea of its function.

To find out, she targeted the patches with her recording electrodes and monitored neuron activity as a monkey viewed pictures of 50 randomly chosen objects — from animals and vehicles to vegetables and houses — each from 24 different angles. The neurons did not respond to faces, but neither did the pattern of firing activity suggest that any other specific category of objects was associated with the network. Instead, the neurons seem to encode general properties of different objects. They seem to register, for example, whether something is spiky like a camera tripod or stubby like a USB stick; animate like a cat or inanimate like a house.

The way that this network processes information has remarkable parallels to how the face-patch network processes faces. Individual cells respond to elements of shape or character, with ramp-shaped tuning. A cell tuned to an object's animacy, for example, might fire slowly for a washing machine and rapidly for a cat. Cells in the more superficial patch tended to respond to similar categories of objects of similar orientation, whereas those in patches deepest in the IT cortex tended to respond to a handful of specific objects, whatever the angle. And Tsao and Bao were able to correctly predict the appearance of any object by looking at firing patterns from just 400 or so neurons.

"We think the entire IT cortex may be using the same organization into networks of connected patches, and the same code for all types of object recognition," says Tsao.

That's an idea that resonates with neuroscientist Georg Keller at the Friedrich Miescher Institute for Biomedical Research in Basel, Switzerland. "It gives hope that such a feature-based coding may operate widely in the brain," he says.

THE HALLUCINATING ENGINE

Now, however, Tsao wants to address the even bigger picture of how the brain captures the entirety of the world, rather than just how it decodes objects. This means understanding not just how visual and other sensory information flowing into the brain is processed, but also how high-level knowledge, which experience has embedded deep in the brain, affects perception. "Think about how we know that a blurry blob on a lake is likely to be a duck," she says.

The brain is not just a sequence of passive sieves fishing out faces, food or ducks, she says, "but a hallucinating engine that is generating a version of reality based on the current best internal model of the world". Her ideas draw on Bayesian inference theory; only by combining perception with high-level knowledge can the brain arrive at the best possible understanding of reality, she says.

One possible mechanism is a long-debated theory called predictive processing, which is currently attracting interest among neuroscientists. Predictive processing holds that the brain operates by predicting how its immediate surroundings will change millisecond by millisecond, and comparing that prediction with the information it receives through the various senses. It uses any mismatch — 'prediction error' — to update its model of the world.

> ***"Think about how we know that a blurry blob on a lake is likely to be a duck."***

To find out what's going on, Tsao wants to learn how the hallucinating engine of the brain is wired. But unsure of which approach will work best, she's trying several simultaneously and recording from ever deeper parts of the brain.

One of her methods involves probing optical illusions, such as the famous face–vase picture. The brain automatically flips between the two perceptions after some seconds of staring at it. By recording single neurons as monkeys stare at the picture, Tsao is trying to identify where and how the flip happens in the brain, and how it resets the internal representation of the world. Another method involves showing a monkey a picture of a familiar face, then morphing it into another familiar face, while recording in the brain. The primate's brain will automatically try to categorize a face as familiar, and at a precise point it will switch its perception of which of the two individuals it is seeing. "Ten years ago, no one would have known where to start investigating these phenomena because we didn't know where faces — or vases — were processed in the brain," says Tsao. Now that both location and code are known, "we can ask questions about exactly what changes as perception shifts".

The approach in non-human primates "has a lot of potential", says Keller, who studies predictive coding in the mouse visual cortex. Mice have a limited internal model of the world, he says, and it is unclear whether results from the mouse will be applicable to people. Although he and others can study predictive coding in the human brain using fMRI and electroencephalograms, such techniques will allow only a superficial inspection. "We won't be able to get at the mechanism, or how it is implemented, in the human like Doris will be able to."

Tsao continues to probe deeper into the brain in search of the sort of beautiful equations that her father inspired her with when she was young. She no longer has to hide her excitement, however. Now, it spreads across her entire face. ■

Alison Abbott *is Nature's Senior European Correspondent.*

1. Tsao, D. Y. *et al. Science* **311**, 670–674 (2006).
2. Moeller, S. *et al. Science* **320**, 1355–1359 (2008).
3. Freiwald, W. A. & Tsao, D. Y. *Science* **330**, 845–851 (2010).
4. Chang, L. & Tsao, D. Y. *Cell* **169**, 1013–1028 (2017).
5. Quian Quiroga, R. *et al. Nature* **435**, 1102–1107 (2005).
6. Popivanov, I. D. *et al. J. Neurosci.* **34**, 95–111 (2014).
7. Kornblith, S. *et al. Neuron* **79**, 766–781 (2013).
8. Lafer-Sousa, R. & Conway, B. R. *Nature Neurosci.* **16**, 1870–1878 (2013).

© 2018 Springer Nature Limited. All rights reserved.



A perpetually dark landscape of lakes and rivers exists underneath Antarctica's thick glacial blanket.

J. T. THOMAS

LIFE BELOW THE ICE

Researchers in Antarctica are drilling through 1,100 metres of ice into a lake that has remained sealed for millennia. Here's what they hope to find.

BY DOUGLAS FOX

At a remote camp just 600 kilometres from the South Pole, the race is on to melt 28,000 kilograms of snow. Within the next two weeks, a team of technicians will use that hot water to melt a hole through 1,100 metres of ice, straight down to the bottom of the Antarctic ice sheet. Their quarry is a hidden lake that has been cut off from the rest of the world for thousands of years. The life they expect to find there inhabits one of the most isolated ecosystems on Earth.

The pool of water, known as Subglacial Lake Mercer, covers 160 square kilometres — twice the size of Manhattan — and might be 10–15 metres deep. Despite temperatures that are likely to stay below 0 °C, the lake doesn't freeze, because of the intense pressure from the ice above. Researchers discovered its ghostly silhouette a little more than a decade ago through satellite observations, but no human has directly observed the lake.

The drillers hope to tap into Mercer sometime around Christmas. Then, a team of researchers from more than a dozen universities will hoist samples of water and mud from its interior. The scientists will also send a skinny, remote-operated vehicle down through the 60-centimetre-wide hole to explore the dark waters with video cameras and grab samples with a claw.

Antarctica conceals more than 400 lakes beneath its ice, and Mercer will be the second that humans have sampled directly. It marks the first time scientists will use a remote vehicle to roam beneath the ice sheet.

Some of the same researchers drilled into a nearby, smaller subglacial pool called Lake Whillans in 2013, and found it teeming with microbes — many more than they had expected in a place cut off from the Sun's energy. This time, they wonder whether the cameras of the submersible might even spy animals in the black waters. "We don't know what's going to be there," says John Priscu, a lake ecologist at Montana State University in Bozeman and leader of the project. "That's what makes it so much fun."

© 2018 Springer Nature Limited. All rights reserved.

Known as SALSA (Subglacial Antarctic Lakes Scientific Access), the expedition is being funded by the US National Science Foundation at a cost of nearly US$4 million. Its aim is to explore the ice-shrouded environment of perpetually sunless rivers, lakes and wetlands that exists in Earth's polar regions and covers an area as big as the United States and Australia combined. This ecosystem is much more isolated than even the deepest ocean trenches.

The subglacial biosphere provides an analogue for habitats deep inside Mars or on the ice-covered moons of Jupiter and Saturn. The scientists leading the project hope that the Lake Mercer ecosystem will shed light on what kind of life can survive in such remote environments.

And the sediments they pull up from the bottom of Lake Mercer could provide clues about how susceptible the ice sheet covering West Antarctica will be to global warming. Cores drilled from the bottom of the Ross Sea nearby suggest that this ice sheet has collapsed dozens of times over the past 6 million years. Lake Mercer is about 800 kilometres inland of those sites in the Ross Sea and could yield important clues about the ice sheet's periodic advances and retreats during previous cold and warm spells, says David Harwood, a bio-stratigrapher at the University of Nebraska–Lincoln who will be present for the drilling. "It would be great to get a record of past changes in the West Antarctic Ice Sheet from that particular location."

HIDDEN HABITAT

Mercer sits within a constellation of nine lakes in West Antarctica that was first discovered in 2006, when satellite altimeter measurements revealed that the ice surface in certain places was periodically rising and falling by up to 10 metres over periods of months[1]. Helen Fricker, a glaciologist at the Scripps Institution of Oceanography in La Jolla, California, realized that they were subglacial lakes filling and emptying, causing the ice overhead to lift and then drop (see 'Land of invisible lakes').

Evidence pulled up from the drilling project at Lake Whillans has spawned a series of discoveries that have shaped the current programme at Lake Mercer, 40 kilometres to the southeast. The water from Lake Whillans teemed with 130,000 microbial cells per millilitre — a population 10–100 times bigger than some researchers expected[2]. Many of the microorganisms obtained their energy by oxidizing ammonium or methane, probably from deposits at the bottom of the lake[3,4]. That was a key insight, because it suggested that this ecosystem — seemingly cut off from the Sun and photosynthesis as an energy source — was still dependent on the outside world in an indirect way.

The researchers who studied Lake Whillans suspect that the ammonium and methane seep up from the lake's muddy floor from the rotting corpses of marine organisms that accumulated during warm periods, millions of years ago, when this region was covered by ocean rather than ice. Evidence of this food source came from Reed Scherer, a micropalaeontologist at Northern Illinois University in DeKalb, who was part of the Whillans project. He found the shells of diatoms (single-celled algae) and the skeletal fragments of sea sponges littered throughout the lake's mud. "There is a marine-resource legacy that the microbes are still tapping into," he says.

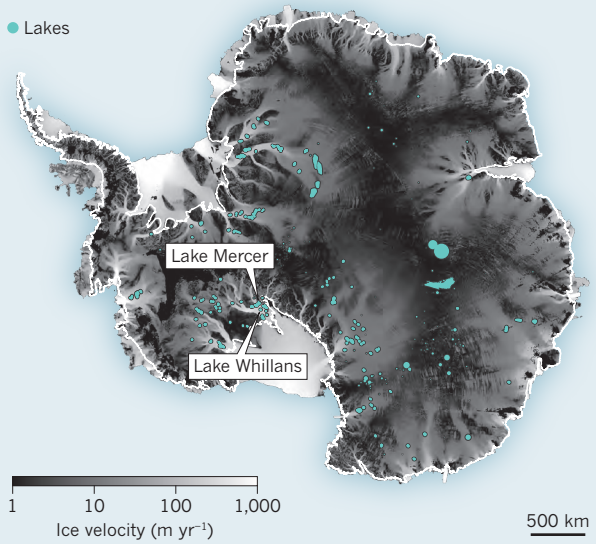
When they drilled into Lake Whillans, researchers thought it had been covered by ice for at least 120,000 years, or possibly up to 400,000 years, coinciding with the last time the West Antarctic Ice Sheet was thought to have melted so dramatically that the lake area had been exposed to the ocean. But in June, Scherer reported evidence that Lake Whillans was connected to the ocean possibly as few as 5,000–10,000 years ago[5].

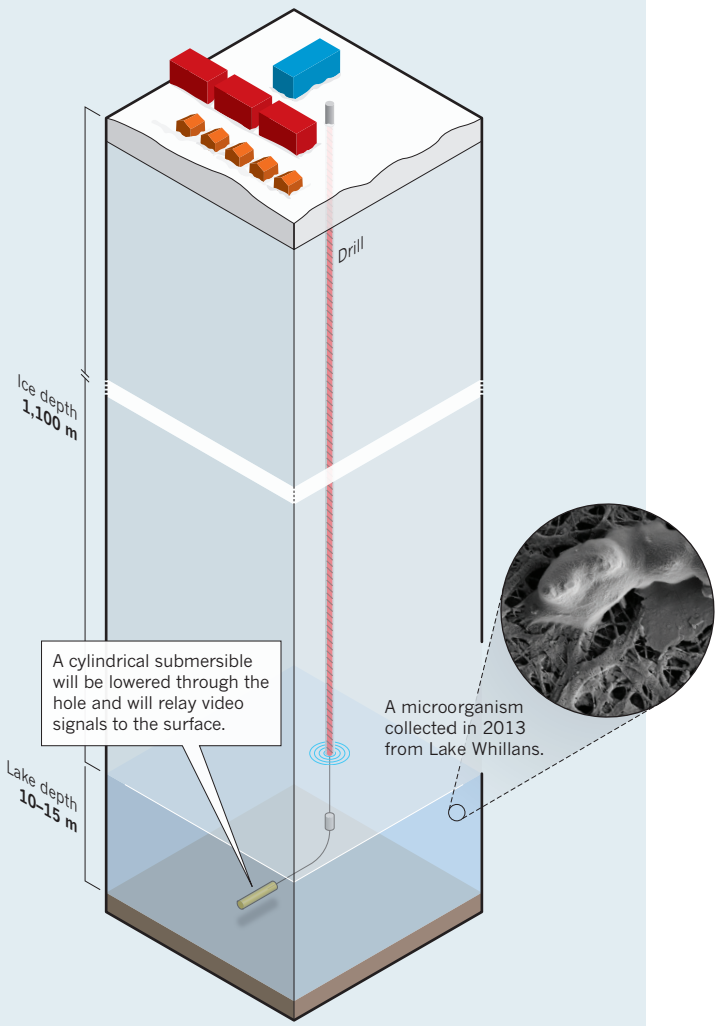
This relatively recent delivery of food has big implications. "It's probably part of the reason we saw such a productive ecosystem" in the lake, says Brent Christner, a microbiologist from the University of Florida in Gainesville who was part of the 2013 expedition and is also involved in the current programme.

When the researchers drill into Lake Mercer, they'll come armed with new instruments to answer some of the questions that emerged from the previous project. Mercer is twice the size of Lake Whillans, and five times deeper — but was probably connected to the ocean at the same time as Whillans, says Scherer. Given the realization that the lakes might

LAND OF INVISIBLE LAKES

Measurements of ice movement and other data have revealed around 400 lakes hidden under Antarctica's ice sheets. Researchers are preparing to explore Lake Mercer by drilling through its 1,100-metre-thick icy cover. Water and sediment samples will reveal the lake's history and provide evidence of any microbial communities living in the dark waters. Lake Mercer lies just 40 kilometres from Lake Whillans, which was explored in 2013.





SOURCES: SALSA; M. R. SIEGFRIED & H. A. FRICKER *ANN. GLACIOL.* **59**, 42–55 (2017).

© 2018 Springer Nature Limited. All rights reserved.



J. T. THOMAS

Researchers pull up sediment from the bottom of Lake Whillans during a drilling project in 2013.

"WE DON'T KNOW WHAT'S GOING TO BE THERE. THAT'S WHAT MAKES IT SO MUCH FUN."

not have been cut off for tens of thousands of years, the team hopes to learn what Lake Mercer's inhabitants are eating. They could subsist on ammonium, methane and other organic compounds from relatively fresh food deposited a few thousand years ago, or they might consume less-easily digested material that is millions of years old. Discovering the microorganisms' diet could help the team to predict how much life might inhabit other subglacial lakes that have been isolated for a lot longer.

It could also hint at how much life, if any, might survive below the surface of Mars — a planet that used to be much more hospitable billions of years ago, when water was present on its surface. Priscu suspects that if life exists on Mars, much of it would be living off carbon that was laid down by photosynthetic organisms, when the planet was wetter.

The team hopes to get a lot of answers about life in Lake Mercer by extracting a core of mud up to 8 metres long from the base of the lake. Brad Rosenheim, a palaeoclimatologist at the University of South Florida in Tampa, will use an advanced technique to analyse the mud for carbon-14, a radioactive isotope formed in the atmosphere that decays to undetectable levels within about 40,000 years. This could reveal how much of the lake-bottom muck was laid down the last time the ocean reached this spot.

Such evidence would provide an estimate of the amount of fresh food that was deposited and the degree to which microbes are now eating this young material, relatively rich in carbon-14, versus gnawing on older carbon. That could help researchers to determine what kind of life might populate other subglacial lakes that have been cut off for an even greater length of time.

Christner hopes to enhance this picture by analysing the core for scraps of DNA left behind by sea sponges, brittle stars, crustaceans or other marine life that arrived with the most recent ocean invasion, several thousand years ago. This could provide further information on what today's resident microorganisms are eating.

Harwood and his colleagues will study the microscopic diatom shells that they expect to find at the bottom of Lake Mercer. By matching up the diatom species found in the core with records of when those species went extinct in other areas, they hope to date the movements of the West Antarctic Ice Sheet.

This evidence could indicate when the ice sheet melted significantly in the past and what the climate was like then, says Harwood. That could help researchers who are trying to forecast when global warming will trigger runaway melting of the West Antarctic Ice Sheet, which in turn could raise sea levels around the globe by several metres.

CREATURES FROM THE DEEP

The SALSA team also hopes to learn whether more-complex organisms, such as animals, might inhabit the subglacial world. In fact, some researchers wonder whether they missed something major when they drilled into Lake Whillans in 2013.

At that time, Whillans's oxygen levels were low — but survivable by a broad range of aquatic animals. And the abundant bacteria in the lake could potentially support microscopic animals such as worms, rotifers or tardigrades. But the researchers were surprised when DNA studies showed no evidence of such creatures[2]. And a video camera that was lowered into the hole for a few minutes did not record any such life.

Then, in 2015, the team discovered a different kind of surprise. They drilled through the ice at another location, 100 kilometres downstream from Lake Whillans, where the ice sheet begins to float on the ocean. Beneath 755 metres of ice, they encountered a sliver of sea water just 10 metres thick. Because the water at that point is more than 600 kilometres from the sunlit edge of the floating ice shelf, researchers did not expect to find complex life. And when they lowered a camera into the hole, the blank images that streamed back confirmed their suspicions — for eight days. Then, they sent down a remote-operated vehicle, and its video cameras soon caught fish, amphipod crustaceans and other animals milling about — in an environment that should not have been able to feed such creatures, because microbial life was scarce.

Priscu now wonders whether Lake Whillans could have also harboured animals that didn't show up in the few minutes of video they captured from their static camera. This time, they are coming to Lake Mercer with better equipment: including the thin submersible that discovered the animals beneath the ice shelf in 2015.

With its three video cameras and claw, it will explore Lake Mercer, venturing up to 100 metres from the borehole and sending images to the surface through a tether. Priscu looks forward to that moment. He thinks that the lake should be capable of sustaining animals.

Peter Doran, a polar scientist at Louisiana State University in Baton Rouge, says that researchers are eager to follow up on the discoveries from Lake Whillans and see what lurks in Mercer. "We need to start building our knowledge, because it turns out that this is a vast ecosystem that's completely unexplored." ■

Douglas Fox *is a journalist in northern California.*

1. Fricker, H. A., Scambos, T., Bindschadler, R. & Padman, L. *Science* **315**, 1544–1548 (2007).
2. Christner, B. C. *et al. Nature* **512**, 310–313 (2014).
3. Michaud, A. B. *et al. Nature Geosci.* **10**, 582–586 (2017).
4. Achberger, A. M. *et al. Front. Microbiol.* **7**, 1457 (2016).
5. Kingslake, J. *et al. Nature* **558**, 430–434 (2018).

© 2018 Springer Nature Limited. All rights reserved.

COMMENT

SOCIETY How water has shaped the history of south Asia **p.186**

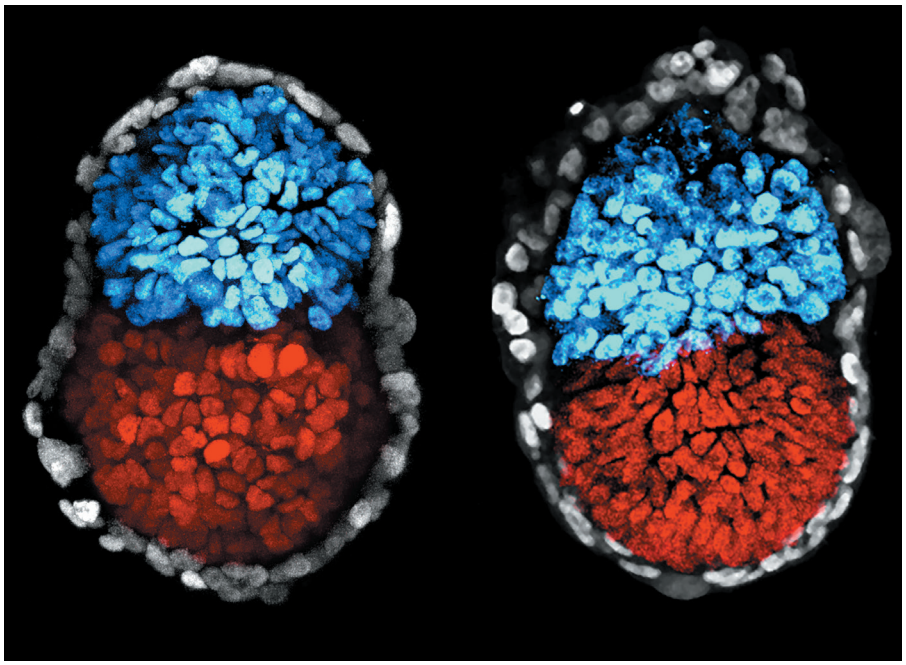


DEXTERITY A chef, a musician and a chemist walk into a conference bar… **p.188**

CRISPR TWINS Gene-therapy regulation offers good model **p.190**

DIVERSITY Redress disparities in NIH funding between institutions **p.190**

B. SOZEN ET AL./NATURE CELL BIOL.



The cells of a 4-day-old artificial embryo (left) resemble those of a 5.5-day-old mouse embryo (right).

Debate ethics of embryo models from stem cells

International discussion must guide research, urge **Nicolas Rivron**, **Martin Pera** and colleagues.

Over the past five years, various studies have shown that mouse and human stem cells can spontaneously organize in a dish into 3D structures that are increasingly similar to mouse[1–5] or human[6–8] embryos. All that is needed is the right number and combination of cells, growth factors and, sometimes, a means of physically confining the cells, such as in microwells[8].

In the past 18 months, researchers have taken a significant step forward, using mouse models. They have incorporated tissues into the models that resemble those that become the yolk sac and placenta. In mammals, these 'extra-embryonic organs'[1,4,5] grow in synergy with the embryo, mediate its implantation and form the interface with the mother.

In short, it now seems feasible that stem cells can be developed into models that are almost indistinguishable from embryos in the lab. Such models can also be transferred into the womb of a mouse[1], where they begin to implant.

These models open up all sorts of possibilities in research. Studying mouse and human embryogenesis in the lab could lead to better infertility treatments or contraceptives, more-effective and safer *in vitro* fertilization (IVF) procedures, the prevention and treatment of developmental disorders and even the creation of organs for people who need a transplant (see 'Why model embryos?').

These models also raise profound ethical questions. What should their legal and ethical status be now, and in the future as they are refined? Do the probable insights these embryo models provide outweigh possible ethical concerns? Because of the potential benefits, is there now a moral imperative to develop this research?

In 2015, various commentators, including four of us (M.P., M.M., G.deW. and W.D.) flagged the potential ethical implications of developing embryo models from stem cells[9]. At the time, investigators had modelled only a short span of development. No precursors of the extra-embryonic tissues had been generated.

Given the pace of progress, we now think that a major international discussion is needed to help guide this research.

NEW AVENUES

So far, biologists have produced four different types of 'embryo model': three in mice[1,3,5] and one using human cells[7] (see 'What's been modelled?' and 'Model systems'). All of the models stop developing after a few days, and the extent to which their gene-expression patterns match those of natural embryos has yet to be rigorously assessed[1,3,5]. Even with these limitations — which are likely to be overcome in the future — stem-cell models open up new avenues for exploring human development and disease.

The first few weeks of development are crucial to the success of a pregnancy and the health of a child[10]. But little is known about how the human embryo forms, implants and develops in the days that follow. Embryos can be observed using ultrasound only after about five weeks. And there are strict regulatory constraints on researchers' ability to manipulate human embryos experimentally.

What is known about this period in human development comes mainly from three lines of research. These are: studies of embryos formed through IVF, including of blastocysts cultured in the laboratory for up to 13 days[11,12]; a small number of archival specimens of human embryos obtained decades ago through surgery and other procedures ▶

© 2018 Springer Nature Limited. All rights reserved.

that would now be considered unethical in most countries (see go.nature.com/2sufgov); and a few comparative studies on closely related primate species, such as cynomolgus monkeys (*Macaca fascicularis*)[13].

Unlike embryos formed through the fusion of a sperm and an egg, model embryos can be generated in large numbers and tweaked, for instance by using gene editing. This means that they can be used in high-throughput genetic tests and drug screens — procedures that generally form the basis of therapeutic discoveries.

Biologists can also use model embryos to uncover basic principles. For instance, it is well known that the placenta supports and instructs the embryo's development. Yet a study this year[1] showed that, in early embryos (blastocysts), the embryo guides the formation and implantation of the future placenta. (That work was led by one of us (N.R.), building on previous observations[14].)

We think that stem-cell-based models could transform medicine in at least five ways (see 'Why model embryos?'). Done properly, studies on embryo models could even obviate some of the ethical conflicts surrounding research on human development: researchers would have less need to study embryos from people or other primates.

FOUR QUESTIONS

Future progress depends on addressing now the ethical and policy issues that could arise.

Ultimately, individual jurisdictions will need to formulate their own policies and regulations, reflecting their values and priorities. However, we urge funding bodies, along with scientific and medical societies, to start an international discussion as a first step. Bioethicists, scientists, clinicians, legal and regulatory specialists, patient advocates and other citizens could offer at least some consensus on an appropriate trajectory for the field.

Two outputs are needed. First, guidelines for researchers; second, a reliable source of information about the current state of the research, its possible trajectory, its potential medical benefits and the key ethical and policy issues it raises. Both guidelines and information should be disseminated to journalists, ethics committees, regulatory bodies and policymakers.

Four questions in particular need attention.

Should embryo models be treated legally and ethically as human embryos, now or in the future?

If the majority view is 'no', biologists could use stem-cell-based models both in basic research and in preclinical applications, unfettered by current legislation or guidelines on human-embryo research. If most stakeholders lean towards 'yes', work involving these models would be permitted in countries that allow the creation of human embryos for research, such as the United Kingdom — subject to the usual ethical and legal restrictions.

Answering this question could require testing whether these entities are capable of developing to term, but such experiments would themselves raise ethical questions. Moreover, the worldwide ban on human reproductive cloning would prevent such a test from being conducted on models formed from induced pluripotent stem cells.

In practice, different models might need to be treated in different ways. For example, it is unlikely that current post-implantation models could ever develop fully into an organism. They mirror only some regions of the embryo, and skip over the developmental stage that normally occurs when it implants in the uterus. Complicating matters, researchers might be able to constrain or enhance the developmental capacity of a particular model using gene editing — such as by incorporating suicide genes that destroy the tissue at a certain point. In other words, what might be considered an embryo could be flipped by genetic means into a non-embryo, and vice versa.

Which research applications involving human embryo models are ethically acceptable?

Most would agree that research into the origin of infertility and genetic diseases, for example, is a worthy goal and probably achievable within current ethical boundaries. Conversely, the use of human embryo models for reproduction is much harder to justify. Such applications are a long way off, but one day it might be feasible to transfer an embryo created from (genetically edited) stem cells to a woman's uterus to treat infertility or circumvent genetic diseases. Most — including the International Society for Stem Cell Research (ISSCR) — rightly argue that it is not morally acceptable to create humans in this way, even setting aside the considerable uncertainty regarding the healthy outcome of a stem-cell-derived pregnancy.

How far should attempts to develop an intact human embryo in a dish be allowed to proceed?

The response to this will depend on the answer to our first question. If human-embryo models are deemed equivalent to human embryos, they will become part of an ongoing debate on the time limits on culturing embryos. In more than 20 countries, it is against the law for researchers to maintain intact human embryos in the laboratory past 14 days of development or beyond the initiation of gastrulation (when three different cell layers appear) — whichever comes first[12].

Does a modelled part of a human embryo have an ethical and legal status similar to that of a complete embryo?

At the moment, the following are not deemed biologically equivalent to a whole embryo: tissues sampled from embryos for diagnostic purposes; embryonic stem cells; and extra-embryonic stem cells. But it is unclear at which point a partial model contains enough

WHY MODEL EMBRYOS?

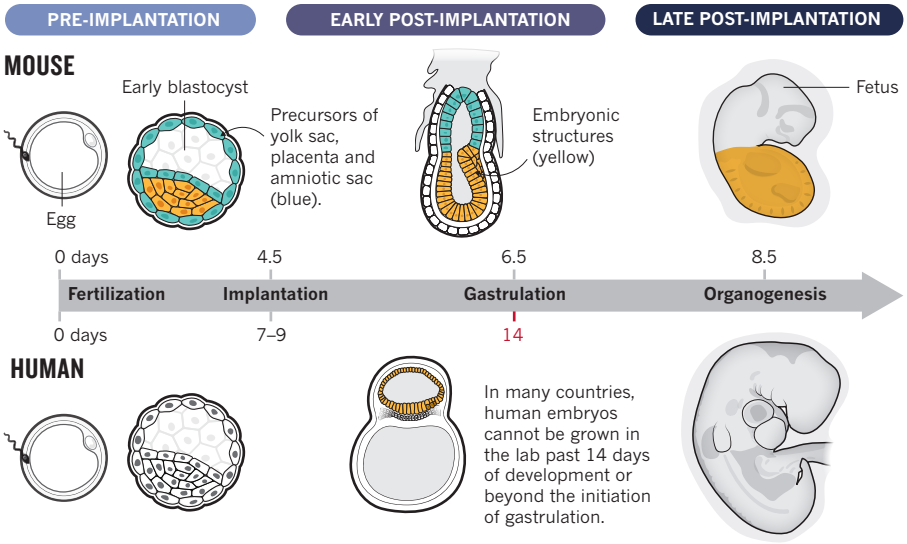
Five ways in which embryo models could improve health

- **Treating infertility.** Embryo models could give researchers a better understanding of implantation and gastrulation, and lead to better infertility treatments. (It is thought that at least 40% of pregnancies fail by 20 weeks, and that 70% of those that fail do so at implantation[15].)
- **Improving IVF.** Only around 20% of IVF procedures result in a birth[16]. Using stem-cell models, researchers could optimize implantation and minimize cellular abnormalities, such as an aberrant number of chromosomes. As well as safeguarding the health of children conceived *in vitro*, this could reduce the number of procedures.
- **Designing new contraceptives.** Embryo-model work could improve drugs that prevent implantation (as the oral contraceptive pill or intrauterine devices do, in part). Women and health professionals need drugs and devices that are easier to use and that have fewer side effects. Family planning is central to sustainable, global development (see go.nature.com/2rdqpvw).
- **Preventing disease.** Subtle cell abnormalities during the first weeks of pregnancy, such as those caused by the use of alcohol or medications, can do damage throughout pregnancy and beyond[17]. They can alter development of the placenta and restrict embryo growth, affecting the baby's birth weight and propensity for chronic diseases (such as those of the heart) decades later[18]. Entities based on stem cells could help researchers to pinpoint the genetic and epigenetic changes involved[18], and assess the effects of diets or drugs[10,16].
- **Creating organs.** Mini brains, livers, kidneys and other organoids made from stem cells are highly simplified. Initiating organ development in an environment as similar as possible to the developing embryo might enable researchers to reliably generate structures that more closely resemble mature, functional organs, for drug screens or even for transplantation. **N.R., M.P. *et al.***

© 2018 Springer Nature Limited. All rights reserved.

WHAT'S BEEN MODELLED?

Using stem cells, researchers have produced entities in the lab that mimic certain structures found in the mouse embryo (coloured parts). Work with human stem cells is less advanced.



MODEL SYSTEMS

How stem cells are used to study embryo development

Mouse stem cells can form 3D structures that resemble the 3.5-day-old mouse embryo (the blastocyst) before it implants in the uterus. These 'blastoids' contain analogues of the three cell lineages thought to form the embryo, placenta and yolk sac. Blastoids implanted into female mice trigger a uterine response. Currently, development stops shortly after implantation[1].

Mouse stem cells can also form entities that are similar to specific regions of the 6.5–8-day-old mouse embryo after it has implanted in the uterus[2–5]. A process called gastrulation, during which the body plan is established, occurs in these regions. These models are termed ETS/X embryo-like structures[4,5] and gastruloids[2,3]. In the first type, some interactions between the embryonic and extra-embryonic tissues are repeated[4,5], the anterior–posterior body axis is laid down and analogues of gastrulating cells are generated[5]. In gastruloids, the three basic germ layers are laid down[2,3] and the precursors of organs develop[3].

Work with human stem cells is less advanced, but is on a similar trajectory. Currently, human stem cells can model aspects of gastrulation[19] and the formation of the beginnings of the amniotic cavity[20]. They can also form 3D asymmetric cysts that model the development of the epiblast–amniotic ectoderm axis[6,7]. As far as we know, this structure arises during the second week, soon after implantation. N.R., M.P. *et al.*

material to ethically represent the whole, so this must also be discussed by regulators.

FOUR RECOMMENDATIONS

These are complex questions, and discussions about all these issues and others will need to be regularly revisited as the field evolves. The pace of progress, however, prompts us to recommend the following.

First, we think that the intention of the research should be considered the key ethical criterion by regulators, rather than surrogate measures of the equivalence between the human embryo and a model. This was the approach taken with cloning. In the late 1990s and early 2000s, many nations prohibited human reproductive cloning, but did not ban the transfer of nuclear material from a somatic cell to an egg to produce a blastocyst and generate lines of stem cells. Here, the key consideration was the intention of the study rather than whether the clone was equivalent to a natural embryo.

Second, we urge regulators to ban the use of stem-cell-based entities for reproductive purposes.

Third, in our view, current stem-cell models that are designed to replicate only a restricted part of development, or that form just a few anatomical structures, should not have the ethical status of embryos.

Finally, we urge any scientist using human stem cells for research to abide by existing guidelines, such as those of the ISSCR. They should send their research proposals to a stem-cell oversight committee or a local independent ethical review board before undertaking any studies, submit their results to peer review and publicize their findings.

As part of ensuring good practice, stem-cell researchers, developmental biologists, human embryologists and others need to reach consensus on what terminology accurately captures the properties of the different models. (Currently, several terms are used interchangeably to describe the various types.) Ideally, terms should reflect the cellular composition and tissue organization of each, and indicate their developmental stage and potential.

Such provisions will help to ensure that this research is conducted ethically. Crucially, the recommendations will also help citizens to understand what researchers are doing, and why. Transparency and effective engagement with the public is essential to ensure that promising avenues for research proceed with due caution, especially given the complexity of the science. ■

Nicolas Rivron *leads the laboratory for synthetic embryology both at the MERLN Institute, Maastricht University, Maastricht; and at the Hubrecht Institute (KNAW), Utrecht, the Netherlands.* **Martin Pera** *is a professor at the Jackson Laboratory, Bar Harbor, Maine, USA.* **Janet Rossant, Alfonso Martinez Arias, Magdalena Zernicka-Goetz, Jianping Fu, Susanne van den Brink, Annelien Bredenoord, Wybo Dondorp, Guido de Wert, Insoo Hyun, Megan Munsie, Rosario Isasi.**
e-mails: n.rivron@maastrichtuniversity.nl; martin.pera@jax.org

1. Rivron, N. C. *et al. Nature* **557**, 106–111 (2018).
2. van den Brink, S. C. *et al. Development* **141**, 4231–4242 (2014).
3. Beccari, L. *et al. Nature* **562**, 272–276 (2018).
4. Harrison, S. E., Sozen, B., Christodoulou, N., Kyprianou, C. & Zernicka-Goetz, M. *Science* **356**, eaal1810 (2017).
5. Sozen, B. *et al. Nature Cell Biol.* **20**, 979–989 (2018).
6. Shao, Y. *et al. Nature Mater.* **16**, 419–425 (2016).
7. Shao, Y. *et al. Nature Commun.* **8**, 208 (2017).
8. Shahbazi, M. N. & Zernicka-Goetz, M. *Nature Cell Biol.* **20**, 878–887 (2018).
9. Pera, M. F. *et al. Nature Meth.* **12**, 917–919 (2015).
10. Bianco-Miotto, T., Craig, J. M., Gasser, Y. P., Van Dijk, S. J. & Ozanne, S. E. *J. Dev. Orig. Health Dis.* **8**, 513–519 (2017).
11. Deglincerti, A. *et al. Nature* **533**, 251–254 (2016).
12. Hyun, I., Wilkerson, A. & Johnston, J. *Nature* **533**, 169–171 (2016).
13. Nakamura, T. *et al. Nature* **537**, 57–62 (2016).
14. Gardner, R. L. & Johnson, M. H. *J. Embryol. Exp. Morphol.* **28**, 279–312 (1972).
15. Norwitz, E. R., Schust, D. J. & Fisher, S. J. *N. Engl. J. Med.* **345**, 1400–1408 (2001).
16. Chen, M. & Heilbronn, L. K. *J. Dev. Orig. Health Dis.* **8**, 388–402 (2017).
17. Cha, J., Sun, X. & Dey, S. K. *Nature Med.* **18**, 1754–1767 (2012).
18. Burton, G. J., Fowden, A. L. & Thornburg, K. L. *Physiol. Rev.* **96**, 1509–1565 (2016).
19. Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. *Nature Meth.* **11**, 847–854 (2014).
20. Taniguchi, K. *et al. Stem Cell Rep.* **5**, 954–962 (2015).

Full affiliations for all authors are listed online (see go.nature.com/2sxyqdt).

© 2018 Springer Nature Limited. All rights reserved.



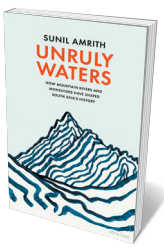
HYDROLOGY

India: a turbulent tale of rivers, floods and monsoons

Philip Ball plunges into an intermeshed human and environmental history.

The Indian novelist Amitav Ghosh remarked on a singular cultural gap in his 2016 book *The Great Derangement*. Environment and climate, he noted, are almost completely ignored in contemporary literary fiction — at precisely the moment they have become major agents of social and political transformation. The same might be said of much twentieth-century historical scholarship: the narratives tend to focus on issues such as urbanization, migration and identity. Now, in his stimulating, urgent *Unruly Waters*, historian Sunil Amrith strives to redress that imbalance.

His focus is the Indian subcontinent, dominated by the monsoon, climate extremes and the great Himalayan rivers, such as the Indus, Ganges and Brahmaputra. It's a tale of drought, flooding, famine, water management and mismanagement — and, looming over all these today, the uncertain consequences of climate change. Water, Amrith shows, infuses Indian culture, influencing political and economic stability, creating inequality and hardship, and acquiring a symbolic charge. (This is evident, for instance, in both Mehboob Khan's 1957 cinematic epic *Mother India* and the environmental activism of writer Arundhati Roy.)

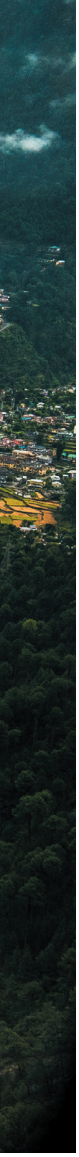


Unruly Waters: How Mountain Rivers and Monsoons Have Shaped South Asia's History
SUNIL AMRITH
Allen Lane (2018)

Much of India's recent history of water resources is a tale of how they were handled under British rule. During the drought-induced famine of 1873–74 in Bihar, in the northeast, the government imported rice from Burma (now Myanmar) and averted crisis. But in 1876–78, the British Raj's response to a drought on the Deccan Plateau and in northwest India was woeful in terms of both preparatory measures and famine relief. Amrith's judgement on this imperial legacy is strikingly relevant today. As he puts it, the Raj effectively undermined local

NATIONAL GEOGRAPHIC CREATIVE/ALAMY

© 2018 Springer Nature Limited. All rights reserved.



The Tehri Dam on the Bhagirathi River, a headstream of the Ganges, in India.

resilience by allowing capitalistic and free-market practices such as taxes on land and produce, and insisting on open markets that drew food away from where it was most needed.

The British colonial governors were convinced that they could engineer Indian modernity. Between 1885 and 1940, the government built a network of irrigation canals in the Punjab to turn "waste" land into cropland, creating prosperous "canal colonies" such as the Chenab settlement. These hydro-engineering schemes altered India's economic landscape while producing a winner-takes-all scenario. "The control of water as well as control of credit concentrated land in fewer hands," further disenfranchising the rural poor, Amrith writes.

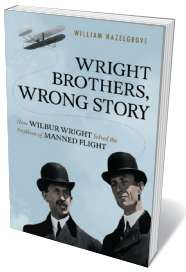
This is just one demonstration of how water management demands holistic thinking. Rivers and waterways do not abide by political borders. The 1947 partition, in which the subcontinent was summarily split into India and Pakistan by the withdrawing British, divided the waters as well as the land and the population. Disputes between the two nations about control of the Indus were among the earliest 'water wars' on the subcontinent, which persist today. Several stem from huge hydro-engineering projects planned for rivers in the Himalayan regions of India, Nepal, Bhutan and Pakistan, which would create 400 dams — roughly 1 every 32 kilometres. China, too, has a major stake in this game, with its plans to dam the Tibetan headwaters of the Brahmaputra.

> *"The control of water as well as control of credit concentrated land in fewer hands."*

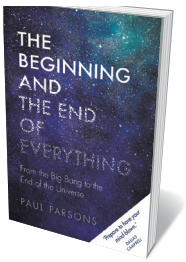
Unruly Waters is an interesting counterpart to studies of water's role in the history of China (my own included). There are as many contrasts as similarities. India does not have quite the stark wet south–dry north climatic division seen in China; nor are its rivers so strategic for trade and conquest. Rather, India's situation shaped its prospects: it is flanked by concave coasts and, after 1869, was accessible from Europe through the Suez Canal. China's climate is also less in thrall to monsoon conditions.

India lacks China's quasi-mythical narrative of civilizational continuity maintained by imperial dynasties. Nor does it have a long history of state-controlled hydro-engineering to exploit the major rivers and to build canals and reservoirs for trade, military transport, water storage and irrigation (A. Janku *Nature* **536**, 28–29; 2016). These two factors — the symbolic and practical value of waterways — are surely connected, even if not in the simplistic and eurocentric idea of "oriental despotism" founded on "hydraulic ▶

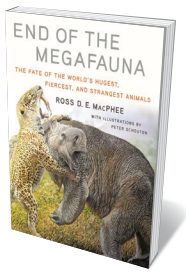
Books in brief



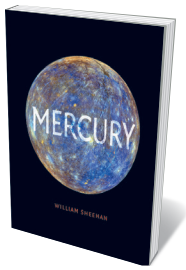
Wright Brothers, Wrong Story
William Hazelgrove PROMETHEUS *(2018)*
In December 1903, the first 12 seconds of controlled, human-powered flight took place near Kitty Hawk, North Carolina. That triumph is engraved in history; less so, the story of Wilbur and Orville Wright, the uber-geeks behind it. In this gripping dual biography, William Hazelgrove argues that theirs was no partnership of equals, as Orville claimed: it was Wilbur who rewrote the science of aeronautics. Hazelgrove delves into their experimental tinkering and family dynamics, but the real story here is that, as he eloquently puts it, one brother was a poet, and the other a scribe.



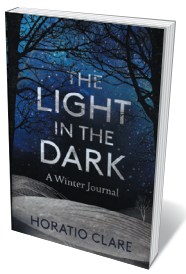
The Beginning and the End of Everything
Paul Parsons MICHAEL O'MARA *(2018)*
If a soup-to-nuts natural history of the Universe appeals, this one is a winner. Paul Parsons, a theoretical cosmologist turned science writer, delivers the oft-told tale with engaging lucidity, from the birth of the Universe 13.8 billion years ago to its putative end in a bang or a whimper aeons hence. As he traverses the phenomena, he interweaves stories of the researchers who discovered them, such as sixth-century Indian astronomer Varahamihira, who first conceptualized a force something like gravity, and the doughty researchers who found gravitational waves in 2015.



End of the Megafauna
Ross D. E. MacPhee W. W. NORTON *(2018)*
Just a few thousand years ago, gargantuan fauna roamed the planet, from the gorilla-sized sloth lemur *Archaeoindris fontoynontii* to the elephant bird *Aepyornis maximus*. What drove the extinction of these species "lost in near time"? Palaeomammalogist Ross MacPhee examines the theories, such as human over-hunting, climate change, emergent infections and food-web disruption; articulates the ongoing debate around them and what that might tell us about today's biodiversity crisis; and takes a look at de-extinction. Packed with evocative artwork by Peter Schouten.



Mercury
William Sheehan REAKTION *(2018)*
Mercury, the Solar System's innermost planet, was spotted in antiquity but remained an enigma until the 1960s. Science historian William Sheehan's portrait of the body (known in ancient Greece as the "scintillating one" for its flicker) reveals it as an airless iron world with an eccentric orbit. He interleaves discoveries, from Johannes Kepler's prediction of a transit of Mercury in the seventeenth century to NASA's MESSENGER probe, which relayed gorgeous images and data (such as the presence of a wealth of volatile compounds on the surface) before crashing on the planet in 2015.



The Light in the Dark
Horatio Clare ELLIOTT & THOMPSON *(2018)*
The leafless gloom of British winters can evoke powerful emotions. Beset by depression during one, nature writer Horatio Clare vowed to track his psychological shifts during the next. His lyrical memoir mines dark realities, from rural crime to seasonal affective disorder and the rising incidence of anxiety among university students. Yet running through all is the understanding that immersion in nature — the "turbulent, colloquial cries of geese", silvered fields and sunlit birches — can help in overcoming the condition, as a growing body of Western and Japanese research suggests. **Barbara Kiser**

© 2018 Springer Nature Limited. All rights reserved.

civilizations" that Marxist historian Karl Wittfogel proposed in the 1950s. The extensive canal building in India occurred mainly under British rule in the nineteenth century, when the use of canals for transport and trade already faced competition from the expanding railway network.

So although Amrith makes a persuasive case that rains, rivers, coasts and seas shaped the history of India as much as they did that of China, they did so in different ways. That is reflected in the fact that mastery of water in India has never been closely linked to a 'heavenly mandate' of state authority, as it has in China. Understanding those distinctions — and perhaps the equally marked differences in water's role in the Middle East — might offer a broader understanding of how history and environment entwine.

Lurking behind these questions is the issue of how far science and technology can help us to understand and manage nature. Modern meteorology can be said to have begun with the British colonial government's efforts to predict the monsoon, although that particular goal is still challenging. (The sensitivity of the monsoon to patterns of global climate such as the El Niño–Southern Oscillation are only now becoming understood.) It's ironic that, just as weather science has started to yield dividends, the impacts of technological advance itself have made it urgent that we develop a longer-term forecast.

In India and China in particular, climate change is complicating the centuries-old struggle with water. Global warming is expected to intensify monsoons, increase weather variability, raise sea levels and melt glacial reservoirs. At the same time, the precipitous modernization and socio-economic development of both countries has exacerbated pollution, over-use and inequalities of access — as potently symbolized in the despoliation of the Ganges.

That's why histories of this kind are needed more than ever. Political, economic and historical discourse cannot just linger on statecraft and strategy, alliances and migrations, trade and war. Increasingly, the environment is central — and its role needs to be understood not through sweeping, Wittfogel-style theses, but with the kind of attention to local detail and nuance that Amrith exhibits.

He is right to assert one general lesson about water management. He writes that it has never been solely a question of technology or science that can be solved within political borders. The unruliness of water means that the business of working with it is "deeply inflected with cultural values, with notions of justice, with ideas and fears about nature and climate". ■

Philip Ball *is a writer based in London, and the author of* The Water Kingdom: A Secret History of China.
e-mail: p.ball@btinternet.com



LAB LIFE

Science in hand: how craft informs lab work

Artists and performers can enrich the physical act of experimentation, explain **Roger Kneebone**, **Claudia Schlegel** and **Alan Spivey**.

Even shaking a sample, rather than stirring it, can change results. Why then, among the many reasons discussed for the reproducibility crisis, does lab practice not get more attention (see *Nature* **533**, 452–454; 2016)?

Most science students enter university with years of screen time under their belts, but very little experimental experience. Indeed, many early-stage PhD students struggle with the transition from predetermined practicals to independent experimentation and design, where the ability to notice tiny departures from the expected might be crucial to discovery.

Some might not have 'good hands'. Moreover, written accounts are notoriously open to interpretation: 'add reagent X dropwise until the solution changes from red to yellow' seethes with potential ambiguity. Laboratory knowing takes place at the intersection between materials, tools and a researcher's body. Its rhythms differ from those of simply absorbing facts.

We — a surgeon, a research nurse and a synthetic chemist — looked beyond science to discover how people steeped in artistic skills might help to close this 'haptic gap', the deficit in skills of touch and object manipulation. We have found that craftspeople and performers can work fruitfully alongside scientists to address some of the challenges. We have also discovered striking similarities between the observational skills of an entomologist and an analytical chemist; the dexterity of a jeweller and a microsurgeon; the bodily awareness of a dancer and a space scientist; and the creative skills of a scientific glassblower, a reconstructive surgeon, a potter and a chef.

For more than 20 years, R.K. has explored this landscape, building a network of experts from apparently unconnected domains to share insights for the lab or operating theatre. In October last year, that multidimensional collaboration led to the Art of Performing Science, a symposium at Imperial College

© 2018 Springer Nature Limited. All rights reserved.





The Art of Performing Science symposium at Imperial College London. L–R: Curator Miranda Lowe works with taxidermist Derek Frampton; letter cutter Phil Surey with plastic surgeon Haz Saddeen; and semiotics scholar Gunther Kress with technician Paul Brown and space scientist Kathrin Altwegg.

FERGUS BURNETT/IMPERIAL COLLEGE LONDON

London — funded by the UK Economic and Social Research Council — that has proved to be a powerful catalyst for further collaboration.

It was an intensely diverse grouping, drawing together more than 60 experts from Britain and the rest of Europe. Here were synthetic chemists, biologists, paediatric surgeons, radiologists, scientific glassblowers and instrument technicians; and social scientists from anthropologists to ethnographers. Here, too, were curators, keepers and conservators from major UK institutions, including London's National Gallery and Victoria and Albert Museum; potters, taxidermists and stonecutters; and performers including musicians and dancers, as well as chefs and even an Olympic rower.

Unexpected parallels in practice emerged. Kathrin Altwegg, leader of the ROSINA instrument programme for the European Space Agency's ROSETTA comet probe, and Imperial technician Paul Brown (who works on the agency's 2020 Solar Orbiter initiative) revealed how space programmes demand close collaboration between experts in their disciplines. Ophthalmic anaesthetist Friedrich Lersch described how his fingers must 'see' layers of the eye to ensure that he finds the right plane for injection. He revealed that his past experience as an apprentice tailor has enabled the interpretation of subtle signals from materials. Conservators Charlotte Hubbard and Isabella Kocum found common ground with taxidermist Derek Frampton and dentist Flora Smyth Zahra; all must manipulate probes and forceps to restore fragile materials. Detailed observation and fine hand–eye coordination are centrally important in the lab, conservation room and artisan's studio.

Better still, 12 months on, some of these encounters have led to practical solutions. Letter-carver Phil Surey and consultant hand surgeon Samantha Gallivan discovered a shared gestural language: both use chisels with hard materials, 'listening' to stone or bone as they make irreversible cuts with focused precision. Carvers, for instance, notice that when stone 'gets tired' through repeated hammering, it is at risk of fragmenting. Gallivan and fellow orthopaedic surgeon Malek Racy are now collaborating with stonecarver Nina Bilbey to find a way of developing such skills early in surgical training.

> ***"Laboratory knowing takes place at the intersection between materials, tools and a researcher's body."***

The parallels between chemistry and cooking were especially striking. Jozef Youssef — chef patron of London-based experimental gastronomical design studio Kitchen Theory — highlighted *mise en place*. This central culinary principle demands that each cook manage their own work space — knives here, ingredients there — to ensure that they can replicate dishes in the demanding setting of high-end restaurants. In labs, similar ways of working are expected but seldom articulated.

R.K., A.S. and Youssef have now launched the Chemical Kitchen, a three-year collaboration between chemistry and culinary students that will start in January 2019. Through a programme of graded tasks, undergraduate chemistry students at Imperial will experience the planning and precision of a professional kitchen. Whether making a soufflé or baking bread, ingredients must be weighed, combined and transformed through heat and pressure, every element rigorously controlled. Would-be chefs, in turn, will practise chemical procedures requiring similar precision, such as distillation or using a Schlenk vacuum line cooled to −50 °C to reproducibly form air-sensitive organometallic compounds. The aim is reciprocal illumination.

Systematic working, close noticing, dexterity, meticulousness and respect for materials and fellow workers all lie at the heart of successful, reproducible science. By examining laboratory 'doing' from unfamiliar perspectives, scientists could shed new light on — and hopefully begin to overcome — the reproducibility crisis. ■

Roger Kneebone, *a trauma surgeon and general practitioner by background, is professor of surgical education and engagement science at Imperial College London.* **Claudia Schlegel** *is head of the Skillslab at the Bern College of Higher Education of Nursing in Switzerland.* **Alan Spivey** *is a synthetic organic chemist trained in Nottingham, Oxford, Geneva and Cambridge, and now a professor at Imperial College London.*
e-mail: r.kneebone@imperial.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

Correspondence

Gene editing: who should decide?

Last month's announcement claiming the birth of the world's first genome-edited babies has sparked a furore over how to regulate this cutting-edge technology (see *Nature* **563**, 607–608; 2018, and *Nature* **564**, 5; 2018). In our view, piling up scientist-led conferences modelled on Asilomar in 1975 (see *Nature* **526**, 293–294; 2015) without any clear consensus is futile.

But lessons can be drawn from another successful case of scientific self-regulation. That is the 1990 Declaration of Inuyama on genetic screening and gene therapy. The Council for International Organizations of Medical Sciences took the lead in calling a conference, made a clear declaration after six days of discussion, and sent it to the World Health Organization, which disseminated it to organizations worldwide. Participants at the conference included biologists, sociologists, psychologists, legal experts, philosophers and religious representatives. Gene therapy could then move from bench to bedside.

International guidelines devised and monitored by scientists could likewise prove useful in regulating genome editing. They should build on previous attempts to do so, for example by the Hinxton Steering Committee in 2015, although the group lacked the necessary diversity (S. Chan *et al. Am. J. Bioeth.* **15**, 42–47; 2015). In return for academic freedom, scientists must regulate themselves — and not just rely on government officials or bioethicists to make such decisions. This regulation would have to involve transparent interaction with citizens.

To restore society's confidence in researchers' professional integrity, rogue germline editing must be stopped by fast and forceful action from genome scientists to lay out transparent rules for gene editing in humans and human embryos. Failure to comply with these rules should incur penalties.

Akira Akabayashi, **Eisuke Nakazawa** *University of Tokyo, Japan.*
Arthur L. Caplan *New York University School of Medicine, New York, USA.*
akirasan-tky@umin.ac.jp

COP24, SDGs: use same stats please

A draft negotiating text for this year's 24th meeting of the Conference of the Parties (COP24) to the United Nations Framework Convention on Climate Change (UNFCCC) aims to strengthen the reporting of nationally determined contributions (see go.nature.com/2arstr1). It also attempts to regulate national statistical processes — which is the mandate of the UN Statistical Commission and of national statistical offices in member countries. In my view, the UNFCCC should instead track progress towards key climate targets by striving to harmonize the data that it requests with those required by other national and international statistical processes.

Under the UN Sustainable Development Goals (SDGs) agenda (see go.nature.com/2apx8ob), countries must now report on 230 or more indicators and apply a significant subset of statistics covered under the UNFCCC. This means that there is a high risk that public money will be used inefficiently. Different ministries will collect similar information under different definitions and global mechanisms might fund the collection of national statistics in developing countries without sufficient coordination.

UN agencies are investigating how to do this better (see go.nature.com/2ap9fve). New rules for climate reporting should capitalize on, not ignore, individual countries' robust statistical systems and data on socio-economic drivers, production and consumption patterns, agriculture, forest activities and land degradation.

Francesco N. Tubiello *Food and Agriculture Organization of the United Nations, Rome, Italy.*
francesco.tubiello@fao.org

Funding is not just for the minority

The US National Institutes of Health (NIH) recognizes the importance of diversity in the country's biomedical research workforce (see, for example, H. A. Valantine and F. S. Collins *Proc. Natl Acad. Sci. USA* **112**, 12240–12242; 2015), but it still has some way to go to achieve it. In my view, the solution lies in redressing the disparities in NIH funding between institutions.

Success rates for grant applications, as well as award sizes, vary with the race, gender, age and institution of applicants and the state from which they are applying. These differences affect where the grant dollars go (see W. P. Wahls *PeerJ* **4**, e1917; 2016) and lead to funding allocations that are heavily skewed in favour of a minority of geographical regions. The top-funded institution alone gets more dollars than do each of 40 entire states; the top 10 institutions each get more dollars than do each of 35 or more states (based on FY2017 values in NIH RePORTER).

Such concentrations of funding provide diminishing marginal returns, even among 'elite' investigators (M. Peifer *Mol. Biol. Cell* **28**, 2935–2940; 2017). A more egalitarian distribution would support more investigators and increase the diversity of scientific approaches, thereby benefiting biomedical research and the taxpayers who support it.

Wayne P. Wahls *University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA.*
wahlswaynep@uams.edu

No trial by media for bullying allegations

Progressive institutions such as the Wellcome Sanger Institute, where we work, implement anti-bullying policies that support independent investigations into whistle-blowers' allegations and empower staff to report concerns. However, media coverage of such disputes can be damaging if it is one-sided (*Nature* **563**, 304–305; 2018). It can render the conclusions of the independent investigation irrelevant to public opinion. A potentially constructive process then adversely affects the reputation of the institution, its staff and their research.

Whistle-blowing allegations cannot be fairly scrutinized in public owing to their complexity, and to legal-privacy protections. Narratives that readily garner the media spotlight risk eclipsing grievances that warrant individualized, rigorous and compassionate redress.

To help ensure that new anti-bullying policies are successful, media coverage needs to be sensitive and balanced, and those who produce it should be aware of its impact.

Grace Collord, **Jyoti Nangalia**, **Luiza Moore** *Wellcome Sanger Institute, Hinxton, UK.*
gc8@sanger.ac.uk

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines and section policies at **http://go.nature.com/cmchno**.

© 2018 Springer Nature Limited. All rights reserved.

CLIMATE SCIENCE

El Niño events set to intensify

After decades of uncertainty, it now seems clear that global warming will enhance both the amplitude and the frequency of climate phenomena known as eastern Pacific El Niño events, with widespread climatic consequences. SEE ARTICLE P.201

YOO-GEUN HAM

During El Niño events, sea surface temperatures (SSTs) in the Pacific Ocean increase. These rising temperatures cause considerable reorganization of atmospheric circulation, resulting in extreme weather events worldwide and strongly affecting ecosystems, human health and the global economy. The reorganization is greater for El Niño events that generate maximum warming in the eastern equatorial Pacific than for those producing maximum warming in the central equatorial Pacific — referred to as EP- and CP-El Niño events, respectively. Despite the huge impact of these phenomena, there has been no consensus on how the SST variability associated with El Niño events will change with global warming[1,2]. But on page 201, Cai *et al.*[3] report robust agreement among climate models that both EP-El Niño SST variability and the frequency of strong EP-El Niño events will increase.

Conventionally, the response of El Niño SST variability to global warming has been investigated in climate models using SSTs at a fixed location. In the case of EP-El Niño events, this location is typically in the eastern equatorial Pacific (the 'Niño3' region: 5° S–5° N, 150°–90° W). Such an approach assumes that all models simulate an EP-El Niño centre — corresponding to the location of peak SST variability — that is the same as the observed centre. Cai and colleagues' breakthrough comes from the realization that this fundamental assumption is invalid. The authors find that the longitude of the simulated centres differs greatly between models, and they examine the response of EP-El Niño SST variability to global warming at the centre of each model.

Another common limitation of climate models is their inability to simulate distinctive CP- and EP-El Niño events[4]. Cai *et al.* show that this limitation reflects a deficiency in simulating asymmetries between CP- and EP-El Niño events, and between these phenomena and their counterpart La Niña events, which are associated with cold SST anomalies (departures from average conditions).

The cold SST anomalies of La Niña events, particularly extreme episodes, tend to occur in the central Pacific. Consequently, in the central Pacific, these anomalies are typically larger than the warm SST anomalies associated with CP-El Niño events — the anomalies are negatively skewed (Fig. 1a). By contrast, in the eastern Pacific, SST anomalies are positively skewed (Fig. 1b). The location of maximum negative SST skewness is the CP-El Niño centre, whereas the location of maximum positive SST skewness is the EP-El Niño centre. As a result, models that more accurately simulate these skewed features produce more-distinctive CP- and EP-El Niño centres.

A technique called empirical orthogonal function (EOF) analysis is often used to study spatial patterns of climate variability and how the amplitude of such patterns changes with time. Data are projected onto these spatial patterns to obtain variables known as principal components, which describe the amplitude of the patterns at each time step. To distinguish between CP- and EP-El Niño centres, at least two principal components representing two distinctive patterns are required[5].

Cai and colleagues obtain these variables

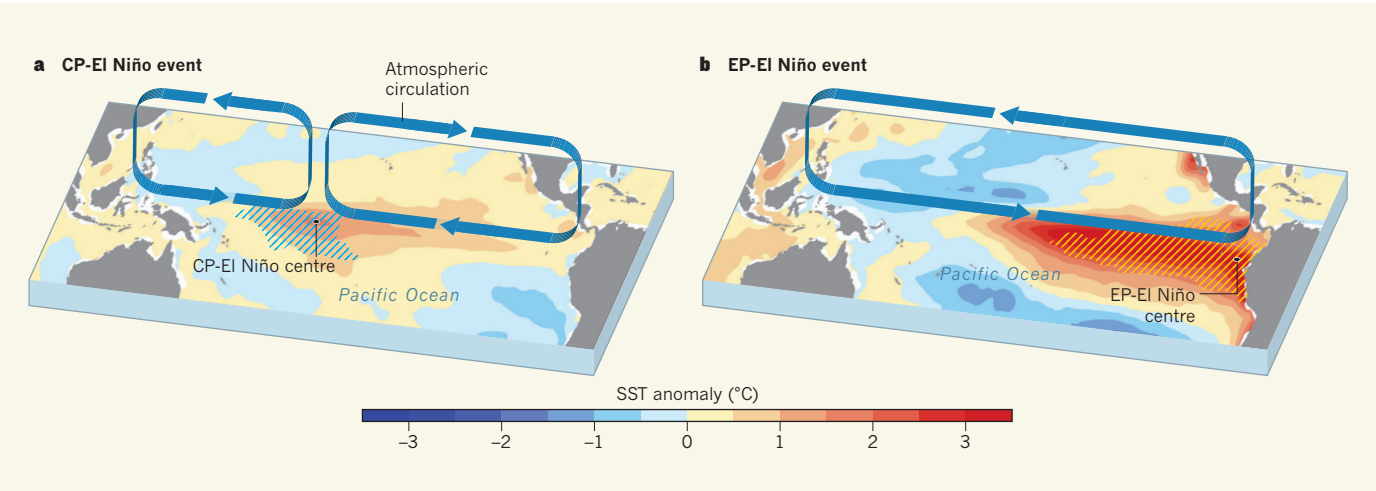


Figure 1 | Two types of El Niño event. El Niño events are associated with changes (anomalies) in sea surface temperatures (SSTs) in the Pacific Ocean. These anomalies result in a reorganization of atmospheric circulation. El Niño events typically have a centre in either the central equatorial or the eastern equatorial Pacific, and are referred to as CP- and EP-El Niño events, respectively. **a**, In the central Pacific, SST anomalies are negatively skewed. Anomalies in the region marked by blue hatching are negatively skewed by more than 0.1 °C from December to February — the season in which El Niño events typically mature — based on data from 1948 to 2015. The anomalies are averaged over the 1990–91, 2002–03, 2004–05 and 2009–10 CP-El Niño events. **b**, In the eastern Pacific, SST anomalies are positively skewed. Anomalies in the region marked by yellow hatching are positively skewed by more than 0.5 °C. The anomalies are averaged over the 1982–83 and 1997–98 EP-El Niño events. Cai *et al.*[3] show that the SST variability associated with EP-El Niño events will increase under global warming. (Data taken from ref. 9.)

© 2018 Springer Nature Limited. All rights reserved.

from an EOF analysis of SST anomalies in the tropical Pacific, which yields two dominant principal-component time series and two associated anomaly patterns[5]. They then use a linear combination of these principle components and patterns to identify an individual EP-El Niño centre for each climate model. Finally, they introduce an EP-El Niño index for each model, which represents the model's EP-El Niño centre and pattern. The authors report that a reasonable consensus emerges: 24 of the 34 available models (71%) predict an increase in EP-El Niño SST variability under a climate-change scenario (known as RCP8.5) that assumes greenhouse-gas emissions will continue to rise steeply throughout the twenty-first century.

However, most of the models underestimate the SST skewness. Cai *et al.* show that nonlinear processes responsible for the negative skewness in the central Pacific are tightly connected to those for the positive skewness in the eastern Pacific, and are represented by a nonlinear relationship between the two principle components. Focusing on 17 models that simulate these nonlinear processes realistically, the authors find an even stronger consensus: 15 of the 17 models (88%) predict a rise in EP-El Niño SST variability under the RCP8.5 emissions scenario.

Cai and colleagues' work shows that, under global-warming conditions, warming occurs more quickly at the surface layer of the ocean than in subsurface layers[6]. This increases the vertical temperature gradient of the ocean, which in turn enhances the dynamical coupling between the atmosphere and the ocean. Consequently, the equatorial ocean–atmosphere system becomes more efficient at converting stochastic fluctuations in winds into a potential EP-El Niño event, leading to increased EP-El Niño variability. The authors' results also indicate that, by a similar mechanism, SST variability in the central Pacific is enhanced (albeit not as strongly as in the eastern Pacific). This translates into an increased frequency of CP-El Niño events and of extreme La Niña events — a conclusion that is consistent with previous studies[7,8].

The authors' finding of increased EP-El Niño variability under global warming represents a milestone in climate research, and will inspire studies of the worldwide impact of future changes in El Niño events. However, the work also raises many questions. For example, why do so many climate models fail to simulate the nonlinear processes associated with the SST skewness? What leads to the large discrepancies in the model simulations? And how sensitive is the reported consensus to future models? Cai and colleagues' results therefore need to be assessed further as other model simulations become available. Nevertheless, the projection of more-frequent and stronger El Niño events must be taken seriously, as we prepare to deal with the consequences of global warming. ■

Yoo-Geun Ham *is in the Department of Oceanography, Chonnam National University, Gwangju, South Korea.*
e-mail: ygham@jnu.ac.kr

1. Collins, M. *et al. Nature Geosci.* **3**, 391–397 (2010).
2. Cai, W. *et al. Nature Clim. Change* **5**, 849–859 (2015).
3. Cai, W. *et al. Nature* **564**, 201–206 (2018).
4. Ham, Y.-G. & Kug, J.-S. *Clim. Dyn.* **39**, 383–398 (2012).
5. Takahashi, K., Montecinos, A., Goubanova, K. & Dewitte, B. *Geophys. Res. Lett.* **38**, L10704 (2011).
6. An, S.-I., Kug, J.-S., Ham, Y.-G. & Kang, I.-S. *J. Clim.* **21**, 3–21 (2008).
7. Cai, W. *et al. Nature Clim. Change* **5**, 132–137 (2015).
8. Yeh, S.-W. *et al. Nature* **461**, 511–514 (2009).
9. Rayner, N. A. *et al. J. Geophys. Res.* **108**, 4407 (2003).

VIROLOGY

How cells hush a viral invader

Viruses can insert a copy of their genetic sequence into a host cell's genome. If the insertion fails, gene expression of unintegrated viral DNA in the nucleus is silenced. How this process occurs has now been uncovered. SEE LETTER P.278

PARINAZ FOZOUNI & MELANIE OTT

Viruses known as retroviruses encode their genetic blueprint in the form of RNA. When these viruses enter a host cell, a viral enzyme generates a DNA version of this RNA sequence that can be permanently integrated into the host-cell genome, despite the host cell's efforts to avoid this outcome. Gene silencing provides a vigorous form of host defence against viral DNA that reaches the nucleus but does not successfully integrate into the genome; however, this mechanism is poorly understood. Such gene silencing can limit gene-therapy approaches that use engineered unintegrated retroviruses. Zhu *et al.*[1] reveal on page 278 that the evolutionarily conserved DNA-binding protein NP220 has a central role in silencing the transcription of unintegrated retroviruses.

When unintegrated retroviral DNA enters the host-cell nucleus it rapidly binds to histone proteins[2], which package DNA, suggesting that retroviral sequences are subject to regulation even before any attempted integration occurs. To identify host factors that might mediate the silencing of unintegrated retroviral DNA, Zhu *et al.* used a gene-editing technique called CRISPR–Cas9. This enabled the authors to eliminate expression of individual genes across the entire genome of human cells grown *in vitro* and to test the effect of this on the silencing of viral genes.

The experimental results led the authors to focus on five host proteins. One was the DNA-binding protein[3] NP220, which is found in the nucleus. Three other proteins of interest — MPP8, TASOR and PPHLN1 — form a multiprotein structure called the HUSH complex that has been previously associated[4] with maintaining the dormant state of integrated HIV, and in contributing to the silencing of integrated

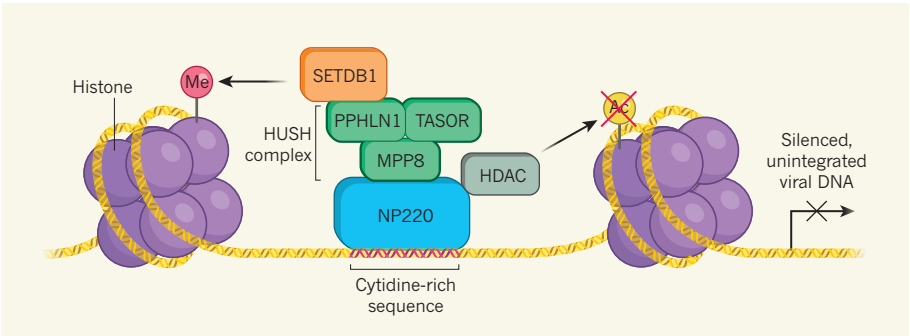


Figure 1 | A protein complex that suppresses viral-gene expression. Zhu *et al.* investigated how human cells silence gene expression of viral DNA that is not integrated into the host-cell genome. The authors report that viral DNA regions that are rich in the DNA building block cytidine are bound by the DNA-binding protein NP220. This protein is associated with the proteins PPHLN1, TASOR and MPP8 that form the HUSH complex (green). Another protein associated with this group is the enzyme SETDB1, which can add methyl groups (Me) to histone proteins that package viral DNA. This type of histone modification usually helps to silence gene expression. The removal of acetyl groups (Ac) on histones also represses gene expression. The authors report that histone deacetylase (HDAC) enzymes that catalyse the removal of such acetyl groups are found in association with NP220.

© 2018 Springer Nature Limited. All rights reserved.

retroviral sequences called retroelements[5,6]. The fifth protein was SETDB1, a methyltransferase enzyme that can modify histone proteins and interact with the HUSH complex to block the transcription of newly integrated retroviral DNA[4–6].

The authors tested the function of each of these proteins in human cells grown *in vitro* that were infected with murine leukaemia virus, a model retrovirus that was used in early attempts at retroviral gene therapy[7]. They found that NP220 binds unintegrated retroviral DNA (Fig. 1) and, on binding, recruits the HUSH complex. This complex then recruits SETDB1 to deposit methyl modifications on histone proteins bound to viral DNA. These histone modifications are associated with the suppression of gene expression. Zhu and colleagues also discovered that NP220 recruits two enzymes, called HDAC1 and HDAC4, from a family of enzymes called histone deacetylases, which catalyse the removal of acetyl groups from histones. A decrease in the level of histone acetylation can repress gene expression.

The authors tested the effect of depleting either NP220 or HDAC1 and HDAC4 in human host cells infected with different types of retrovirus. For example, they tested HIV, which is from a different genus of retroviruses from that of murine leukaemia virus, and found that the depletion of these components caused an increase in gene expression from the unintegrated retrovirus. However, the depletion of HUSH-complex components or SETDB1 did not cause increased expression of the unintegrated viral sequence in this scenario. A similar pattern of results was obtained when the authors conducted the same type of study using Mason–Pfizer monkey virus, which belongs to yet another genus of retroviruses. This suggests that the role of NP220 is evolutionarily conserved, but that the proteins that might act with NP220 to silence viral gene expression can vary depending on the retrovirus.

NP220 can bind sequences in double-stranded DNA that are rich in the nucleoside cytidine[3]. Zhu and colleagues report that there are sequences rich in the DNA building block cytidine in repeat sequences called long terminal repeats (LTRs) at the ends of murine leukaemia virus, HIV and Mason–Pfizer monkey virus sequences, and that these cytidine-rich LTR sequences can serve as binding sites for NP220. When the authors tested the effect of depleting NP220 in human cells infected with Rous sarcoma virus (from another retrovirus genus), which has cytidine-poor LTRs, this depletion did not affect the gene expression of unintegrated viral DNA. This suggests that NP220 needs cytidine-rich DNA sequences to bind and silence viral genes. However, the NP220-independent silencing of gene expression of unintegrated Rous sarcoma viral DNA was found to be SETDB1 dependent because the deletion of the gene encoding SETDB1 led to increased gene expression of the viral DNA.

It would be interesting to learn more about how unintegrated Rous sarcoma virus is silenced through this NP220-independent mechanism. Together, these studies reveal that NP220 or its interacting partners can silence unintegrated DNA from a range of retroviruses.

The HUSH proteins and SETDB1 are involved[4–6] in the transcriptional silencing of integrated retroviral elements. Whether NP220 also acts to silence the expression of integrated retroviral DNA remains to be determined. This is a possibility, given that the cytidine-rich binding motifs for NP220 are preserved in viral LTRs after their integration into a host genome. Clearly, there is more to investigate about the silencing mechanism uncovered by Zhu and colleagues.

The high level of evolutionary conservation of NP220 proteins across vertebrate species is noteworthy. The origin of retroviruses is at least as ancient as that of vertebrates[8]. Perhaps NP220 represents an ancestral defence system to tackle invading retroviruses. But, in turn, some retroviruses could have evolved mechanisms to evade such gene silencing, which might explain why Rous sarcoma virus has lost cytidine-rich LTRs that would provide NP220 binding sites. Zhu and colleagues' study might inspire new ways to prevent the silencing of foreign DNA for use in gene-therapy applications. Importantly, it also provides clues to the evolutionary arms race between humans and retroviruses, and offers insights into a mechanism that might offer protection against disease-causing viruses. ■

Parinaz Fozouni *and* **Melanie Ott** *are at the Gladstone Institutes, San Francisco, and in the Department of Medicine, University of California, San Francisco, California 94143, USA.*
e-mail: melanie.ott@gladstone.ucsf.edu

1. Zhu, Y., Wang, G. Z., Cingöz, O. & Goff, S. P. *Nature* **564**, 278–282 (2018).
2. Wang, G. Z., Wang, Y. & Goff, S. P. *Cell Host Microbe* **20**, 798–809 (2016).
3. Inagaki, H. *et al. J. Biol. Chem.* **271**, 12525–12531 (1996).
4. Tchasovnikarova, I. A. *et al. Science* **348**, 1481–1485 (2015).
5. Robbez-Masson, L. *et al. Genome Res.* **28**, 836–845 (2018).
6. Liu, N. *et al. Nature* **553**, 228–232 (2018).
7. Thomas C. E., Ehrhardt, A., & Kay, M. A. *Nature Rev. Genet.* **4**, 346–358 (2003).
8. Aiewsakun, P. & Katzourakis, A. *Nature Commun.* **8**, 13954 (2017).

This article was published online on 28 November 2018.

IMMUNOTHERAPY

Adrenaline fuels a cytokine storm

Attempts to boost the body's antitumour immune responses can trigger a harmful inflammatory reaction called a cytokine storm. New insights into the mechanisms involved might help to prevent this problem. SEE LETTER P.273

STANLEY R. RIDDELL

Many newly developed, potent cancer therapies aim to harness an immune response to target tumours[1]. However, a common problem with such immunotherapy approaches is the development of a severe inflammatory response called a cytokine storm[2,3], in which levels of proteins called cytokines become abnormally high. This results in fever, low blood pressure, heart problems and, in some cases, organ failure and death. There is therefore great interest in understanding the underlying mechanisms to develop ways of preventing cytokine storms without altering the effectiveness of anti-cancer treatments. On page 273, Staedtke *et al.*[4] reveal that the protein ANP can block cytokine storms, and they uncover a self-amplifying production loop in immune cells that generates a class of molecule called catecholamines, which includes the hormone adrenaline (also known as epinephrine). They report that this catecholamine production helps to initiate and maintain cytokine storms.

When immune cells recognize a molecule that indicates a possible threat, they release cytokines that promote inflammation and orchestrate host defence[5]. One antitumour treatment that can trigger a cytokine storm uses a bacterium called *Clostridium novyi*-NT, which tracks to the low-oxygen environments found in certain tumours and releases spores that cause tumour-cell death[6]. Determining the correct bacterial dosage is difficult, and mice that have large tumours and receive a high dose of *C. novyi*-NT often develop a fatal cytokine storm that cannot be prevented by using inhibitor molecules to block the actions of cytokines or their receptors[6].

To determine whether some known anti-inflammatory proteins could block a cytokine storm, Staedtke and colleagues engineered *C. novyi*-NT to secrete anti-inflammatory proteins and tested whether any of these bacteria could treat tumours effectively without

© 2018 Springer Nature Limited. All rights reserved.

causing severe toxicity owing to high cytokine levels. Their experiments revealed that ANP can dampen a cytokine storm. Mice treated with ANP-expressing *C. novyi*-NT had lower levels of proinflammatory molecules, including cytokines, in their bloodstream, and lower levels of organ infiltration by immune cells called myeloid cells that are associated with a cytokine storm, compared with mice given *C. novyi*-NT that had not been engineered to express ANP.

To determine how ANP decreased cytokine storms in their model system, Staedtke and colleagues characterized the differences between mice treated with the ANP-expressing *C. novyi*-NT and those that received non-engineered bacteria. This revealed that the dampened immune response linked to ANP was accompanied by a decrease in the level of catecholamines in the animals' bloodstream. Catecholamines such as adrenaline are best known for their role as part of the 'fight or flight' response to acute stress, in which they are released by certain neurons and by the adrenal gland. The idea that catecholamines might act to promote cytokine storms seems counter-intuitive, given that molecules of this class are used routinely to treat the low blood pressure associated with cytokine storms. However, it was known[7] that two types of immune cell — macrophages and neutrophils — produce catecholamines in response to inflammatory stimuli such as the molecule lipopolysaccharide (LPS), which is a hallmark of many types of bacterial infection.

To investigate whether catecholamines might have a key role in driving strong inflammatory responses, Staedtke *et al.* gave mice LPS and also gave a subset of these animals adrenaline. The animals that received adrenaline and LPS had higher cytokine levels and mortality than did those that received only LPS. Conversely, when the authors gave LPS to mice whose macrophages had been engineered to lack an enzyme called tyrosine hydroxylase (which is needed to make catecholamines), the animals had better survival rates and lower levels of cytokines and catecholamines than did LPS-treated mice that had macrophages with intact tyrosine hydroxylase. When the authors treated mice with a drug that blocks a receptor for catecholamines called the α1 adrenergic receptor, this interference with catecholamine signalling reduced inflammation when the mice were treated with LPS, compared with LPS-treated mice that did not receive the drug.

The authors also demonstrated the importance of catecholamines in initiating cytokine storms induced by bacteria in a different model system of severe bacterial infection. In both settings, the authors found that animals given metyrosine, a drug that inhibits tyrosine hydroxylase, had lower catecholamine and cytokine levels and increased survival rates compared with mice that did not receive the inhibitor.

The next key question was whether catecholamine release has a function in cytokine storms that arise from immune-cell activation for reasons other than encounters with a bacterium. Immune cells called T cells that have been triggered to launch an immune response can also make catecholamines[8]. Certain immunotherapy approaches aim to generate such activated T cells by the administration of antibodies that can activate T cells or by the introduction of engineered T cells (called chimaeric antigen receptor (CAR) T cells) designed to target tumour cells. These approaches can cause a cytokine storm[9,10]. To test whether catecholamines might have a role in such cytokine storms, the authors administered a T-cell-activating antibody to a group of mice, and treated a subset of the mice with metyrosine. The animals that received the inhibitor had improved survival and lower cytokine levels than the mice that did not receive metyrosine.

The authors then studied human CAR-T cells that were grown *in vitro* together with the type of blood cancer cells that activate them. The medium from these cell cultures contained catecholamines and cytokines, and the levels of these molecules increased if adrenaline was added to the culture, providing support for a model of a self-amplifying response driving their production.

The authors went on to give CAR-T cells to mice carrying tumours. A subset of the mice were given ANP or metyrosine before receiving the CAR-T cells, and these animals' cytokine levels were lower than were those of the mice that received only CAR-T cells. However, this difference did not affect the efficiency of the antitumour treatment, suggesting that toxicity due to cytokines is independent of the antitumour effects of this treatment.

Staedtke and colleagues provide compelling evidence for a self-amplifying circuit of catecholamine release by immune cells in the initiation of a cytokine storm (Fig. 1). However, determining the details of this circuitry will require additional studies. For example, how immune-cell activation drives an increase in catecholamine levels and how catecholamines boost cytokine production is unknown and should be investigated. Another mystery is which types of adrenergic receptor are crucial for the effects of catecholamines on cytokine levels in humans. ANP has anti-inflammatory properties[11], but how it inhibits catecholamine production is another key unanswered question that deserves future study.

The authors' findings might lead to new strategies to tackle cytokine storms during immunotherapy. Mouse models of CAR-T cell immunotherapy indicate that the activation of myeloid cells has a key role in driving cytokine storms — pre-emptive blockade of the action of certain cytokines or their receptors by antibodies or other approaches can effectively prevent the storms[12,13]. However, Staedtke *et al.* now also identify a central role for catecholamine production in the generation of cytokine

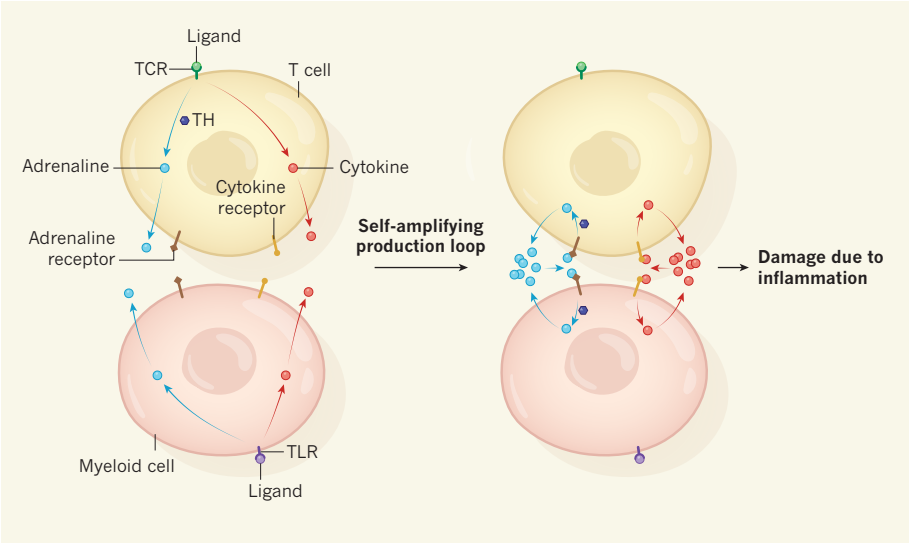


Figure 1 | The pathways driving the harmful inflammatory response called a cytokine storm. Immunotherapy treatments try to boost the responses of immune cells such as T cells against tumours. However, toxicity can occur if immunotherapy triggers a cytokine storm, in which the levels of immune-signalling proteins called cytokines become abnormally high and cause tissue damage. Staedtke *et al.*[4] report studies, using mice and human cells, which reveal that a class of molecule called catecholamines, which includes the hormone adrenaline, has a key role in driving cytokine storms. A T cell can become activated if a ligand molecule binds to the T-cell receptor (TCR), and an immune cell called a myeloid cell can be activated if a ligand binds to its Toll-like receptor (TLR). The activation of these cells leads to the production and release of cytokines, as well as to the production and release of adrenaline. The enzyme tyrosine hydroxylase (TH) catalyses the first step needed for adrenaline production. The authors' work supports a model suggesting that when adrenaline and cytokines bind their respective receptors on immune cells, this increases the production of these molecules through a self-amplifying loop and causes a cytokine storm. Staedtke and colleagues report that if tyrosine hydroxylase is inhibited by the drug metyrosine, it can help to limit cytokine storms (not shown).

© 2018 Springer Nature Limited. All rights reserved.

storms, and show that ANP and metyrosine, which are approved for use in the clinic in other contexts, might be effective in preventing this complication. It is generally assumed that the production of cytokines and their role in the activation of immune cells contributes to the efficiency of antitumour immune responses[14]. To ensure that antitumour effects are not diminished, it will be necessary to proceed cautiously when testing whether targeting catecholamine synthesis can reduce cytokine storms in a clinical setting. ■

Stanley R. Riddell *is at the Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.*
e-mail: sriddell@fredhutch.org

1. Rosenberg, S. A. *Nature* **411**, 380–384 (2001).
2. Gangadhar, T. C. & Vonderheide, R. H. *Nature Rev. Clin. Oncol.* **11**, 91–99 (2014).
3. Hay, K. A. *et al. Blood* **130**, 2295–2306 (2017).
4. Staedtke, V. *et al. Nature* **564**, 273–277 (2018).
5. Medzhitov, R. *Nature* **454**, 428–435 (2008).
6. Roberts, N. J. *et al. Sci. Transl. Med.* **6**, 249ra111 (2014).
7. Flierl, M. A. *et al. Nature* **449**, 721–725 (2007).
8. Bergquist, J., Tarkowski, A., Ekman, R. & Ewing, A. *Proc. Natl Acad. Sci. USA* **91**, 12912–12916 (1994).
9. Chatenoud, L. *et al. Transplantation* **49**, 697–702 (1990).
10. Sadelain, M., Riviere, I. & Riddell, S. *Nature* **545**, 423–431 (2017).
11. Ladetzki-Baehs, K. *et al. Endocrinology* **148**, 332–336 (2007).
12. Norelli, M. *et al. Nature Med.* **24**, 739–748 (2018).
13. Giavridis, T. *et al. Nature Med.* **24**, 731–738 (2018).
14. Kammertoens, T. *et al. Nature* **545**, 98–102 (2017).

The author declares competing financial interests: see go.nature.com/2aycsnd for details.

EVOLUTION

Finches choose parent lookalikes as mates

A preference for mating with similar individuals can have a key role in speciation. Research on Darwin's finches suggests that individuals might use the likeness of their parents as a guide for choosing mates.

LEWIS G. SPURGIN & TRACEY CHAPMAN

New species form when groups of individuals in a population become reproductively isolated and can no longer mate with each other to produce living, healthy offspring. For decades, evolutionary biologists have sought to understand the links between an individual's choice of mate and reproductive isolation between populations and species[1]. Writing in *Proceedings of the National Academy of Sciences*, Grant and Grant[2] provide evidence suggesting that two species of Darwin's finch learn features of their parents early in life and use this knowledge to inform their choice of mate in adulthood, a process known as sexual imprinting. Their study raises fascinating questions about the roles of learning and genetics in mate choice, and how matings between similar individuals (assortative mating) drive the evolution of new species.

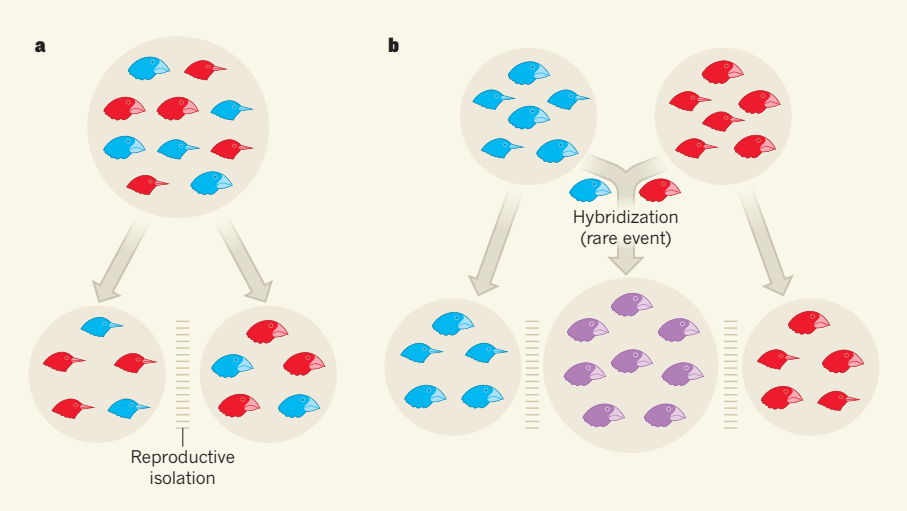


Figure 1 | Potential contributions of sexual imprinting to speciation. Grant and Grant[2] report that, in Darwin's finches, sexual imprinting and assortative mating — learning parental features early in life and using this to choose a mate — can reduce the likelihood of dissimilar individuals mating with each other. Reproductive isolation caused by these and other factors might contribute to the evolution of new species in two ways. **a**, Speciation by fission involves the splitting of one species into two. In this example, imprinted mating preferences for beak size promote the formation of two new species with either small or large beaks. **b**, In speciation by fusion, a rare hybridization event between individuals of different species, followed by an imprinted preference to reproduce with similar individuals, promotes the formation of a new species alongside the two original ones.

Darwin's finches live in the Galapagos archipelago. They are an iconic group of approximately 15 bird species that have contributed hugely to our understanding of natural selection and speciation[3–5]. Previous work has shown that the cultural inheritance of song can promote reproductive isolation between different species of Darwin's finch[6]. However, it was not known whether sexual imprinting based on morphological features such as body size and beak characteristics could similarly promote reproductive isolation, or play a part in the rare cases of mating between species that produce hybrid individuals.

If sexual imprinting is key in directing mate choice, then individuals should choose mates that resemble their parents, and also themselves. In addition, if sexual imprinting contributes to matings between species, then the parents of the hybrid individuals that result from such matings should more closely resemble the other species than their own. To test these hypotheses, Grant and Grant analysed 22 years of data on body size, beak size and beak shape in two finch species — *Geospiza fortis* and *Geospiza scandens* — living on the same island.

Grant and Grant found significant positive associations between certain features of the birds' chosen mates and those of their parents. For *G. fortis*, the body size of the chosen mate was strongly correlated with the body size of the chooser's father, but weakly correlated with that of the chooser's mother. The researchers did not explicitly test whether this imprinting was stronger in male or female offspring. For *G. scandens*, the beak length of male mates chosen by females was significantly associated with the beak shape and length of the female's father, although there were no other significant associations. Grant and Grant suggest that these imprinting patterns can promote assortative mating by body size and beak length in both species.

For matings between species, the results were less straightforward. In some cases, hybridizing individuals or their fathers from a given species were similar in size and shape to individuals from the other species. Although such results are intriguing, hybridization events are rare and sample sizes are small, so

© 2018 Springer Nature Limited. All rights reserved.

further research is needed to confirm whether they can be generalized to other species.

Grant and Grant frame their study as a test of sexual imprinting, but acknowledge that the correlations they observed could also arise if mating preferences were inherited genetically. There has been increasing support for the idea that sexual imprinting can reinforce reproductive isolation in birds and other species[5]. However, evidence for the existence of genetically inherited mating preferences in birds is limited. It is not yet clear whether learnt behaviour has a greater effect on mate choice than does genetic inheritance, or whether these inherited effects have been under-studied. Disentangling the roles of inherited and learnt mate preferences, and their consequences for speciation, is a key challenge for the future[7].

The most powerful tests for identifying sexual imprinting use an experimental 'cross-fostering' approach, in which offspring are swapped early in life and reared by unrelated individuals of the same or a different species[8]. There is also increasing interest in directly quantifying the genetic basis of mate choice using DNA sequencing[9]. We anticipate that future studies will combine experimental and genetic approaches to understand when and why learnt and inherited mating preferences evolve.

Grant and Grant's findings hint that sexual imprinting might have different effects in males and females and across different species. In a cross-fostering experiment in wild mice published last year, the strength of sexual imprinting differed substantially between two species[8]. Furthermore, in the mouse species in which imprinting was weaker, only males showed signs of imprinting. Why imprinting might be weaker and inherited preferences stronger in females of some species is not clear. This could occur if matings between species require more investment in time or effort from females than males, or if mate-choice patterns are influenced by differences in the extent of parental care or the social environment. The effects of inherited and learnt mate preferences are likely to be complex, and studies of a broad range of biological systems might be required to uncover their relative roles in nature.

The work by Grant and Grant links individual variation in mating preferences in Darwin's finches to the evolution of reproductive isolation, which is central to speciation. Sexual imprinting could have a role both in the 'classic' model of speciation, in which one species separates into two, and in the rarer process of speciation through hybridization, in which two different species mix to create a new one (Fig. 1). The possibility that Darwin's finches show sexual imprinting should encourage further experimental tests in other species to determine the role of imprinting in natural populations. Understanding how mating preferences evolve will shed light on the processes shaping past, present and future biodiversity. ■

Lewis G. Spurgin *and* **Tracey Chapman** *are in the School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.*
e-mails: l.spurgin@uea.ac.uk; tracey.chapman@uea.ac.uk

1. Irwin, D. E. & Price, T. *Heredity* **82**, 347–354 (1999).
2. Grant, P. R. & Grant, B. R. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1813662115 (2018).
3. Grant, P. R. & Grant, B. R. *How and Why Species Multiply: The Radiation of Darwin's Finches* (Princeton Univ. Press, 2011).
4. Lamichhaney, S. *et al. Science* **359**, 224–228 (2018).
5. Verzijden, M. N. *et al. Trends Ecol. Evol.* **27**, 511–519 (2012).
6. Grant, B. R. & Grant, P. R. *Evolution* **50**, 2471–2487 (1996).
7. Yeh, D. J., Boughman, J. W., Sætre, G.-P. & Servedio, M. R. *Evolution* https://doi.org/10.1111/evo.13500 (2018).
8. Delaney, E. K. & Hoekstra, H. E. *Evolution* **72**, 274–287 (2018).
9. Merrill, R. M. *et al.* Preprint at *bioRxiv* https://doi.org/10.1101/282301 (2018).

This article was published online on 26 November 2018.

ORGANIC CHEMISTRY

An exciting tool for asymmetric synthesis

A catalytic process driven by visible light converts a mixture of mirror-image isomers of compounds called allenes to a single mirror-image isomer — opening up avenues of research for synthetic chemistry. SEE LETTER P.240

CHENG YANG & YOSHIHISA INOUE

Molecules can exhibit a handedness, known as chirality. This is crucial to many aspects of chemistry and biology because the mirror-image isomers (enantiomers) of a chiral molecule can have distinctly different properties, reactivities and chemical or biological functions. For example, nature often uses just one enantiomer of a family of molecules as building blocks to construct sophisticated structures such as DNA, and in other biological processes. The development of methods for synthesizing chiral molecules asymmetrically — predominantly as a single enantiomer — is therefore one of the most important goals in organic and medicinal chemistry. On page 240, Hölzl-Hobmeier *et al.*[1] report an approach that can also be used to achieve a seemingly impossible task in asymmetric synthesis: the light-induced, catalytic and apparently irreversible formation of single enantiomers of molecules called allenes from a one-to-one mixture of enantiomers (a racemic mixture).

One modern approach to asymmetric synthesis is to use light to induce the formation of a particular enantiomer of a molecule, a strategy called photochirogenesis. Often complementary to conventional methods of asymmetric synthesis, photochirogenesis is useful for making single enantiomers of molecules

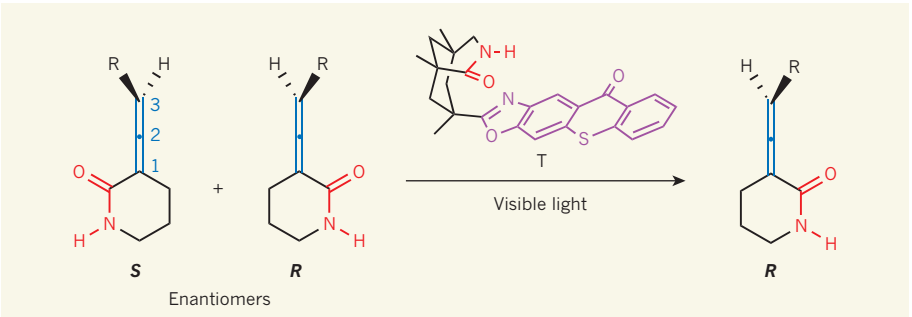


Figure 1 | A light-activated deracemization process. In allene molecules, one carbon atom (designated C2) forms double bonds to its neighbouring carbon atoms (C1 and C3; allene structure shown in blue). If two different groups are attached to each of C1 and C3, two mirror-image isomers (enantiomers) of the allene can form. Hölzl-Hobmeier *et al.*[1] report a light-driven process known as a deracemization, in which a one-to-one mixture of allene enantiomers (*S* and *R*) is converted into just the *R*-enantiomer. The reaction requires a catalyst called an enantiomeric photosensitizing template (T), which contains a photosensitizer group (purple) that allows it to absorb visible light and transfer the energy to the allene. The motifs in red allow T to form a complex with *S* and *R*, as needed for the deracemization. R within the enantiomers represents various chemical groups; the solid wedge and the broken wedge represent bonds that project above and below the plane of the page, respectively.

© 2018 Springer Nature Limited. All rights reserved.



50 Years Ago

The Acting Administrator of the National Aeronautics and Space Administration … shows every sign of confidence that two Americans will tramp about on the surface of the Moon some time next year. The last flight of a team of three astronauts in October seems enormously to have cheered up those responsible for the Apollo programme. Even the accident this weekend which destroyed one of the machines being used to test the rocket system for descending the last few hundred feet to the surface of the Moon seems to have left them unmoved … Plans are now well advanced for the journey around the Moon of the Apollo 8 spacecraft, now assembled at Cape Kennedy, due to begin some time during the week of December 21.
From *Nature* 14 December 1968

100 Years Ago

A writer in the *Times*, directing attention to the fact that a large number of Royal Air Force officers will shortly be demobilised, suggests that they might profitably be employed in making an aerial photographic survey of the British Isles. He believes that this would prove useful to surveyors, architects, engineers, and others. While fully endorsing this writer's opinion that it would be unfortunate to lose the expert services of these flying officers … we cannot agree that a series of aerial photographs could be of great service to surveyors and engineers. Such photographs show the landscape from a new point of view, but they naturally lack the accuracy of carefully drawn topographical maps. On the other hand, such a survey might be of considerable value in the progress of flying for commercial and other purposes. Many attempts have been made to devise suitable maps for airmen, but even the best available leave much room for improvement.
From *Nature* 12 December 1918

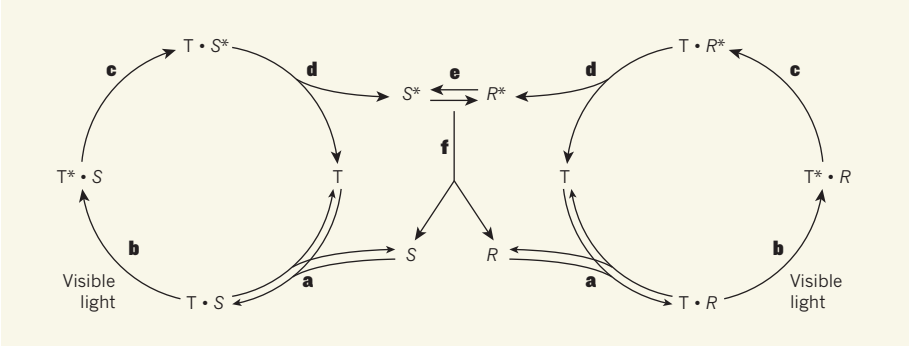


Figure 2 | A plausible mechanism for light-activated deracemization. **a**, T reversibly forms a complex with either *S* or *R*. **b**, Visible light excites T to generate an excited 'triplet' state (T*). **c**, Energy from T* is transferred to *S* or *R*, which enter triplet states (*S** or *R**, respectively), and T returns to its ground state. **d**, The T·*S** (or T·*R**) complex comes apart, releasing T and *S** (or *R**). **e**, **f**, The excited molecules can interconvert (**e**), and eventually relax to their ground states (**f**). However, thermodynamic factors in step **a** and kinetic factors in step **c** ensure that most of the molecules end up as *R*, rather than *S*.

that are difficult or tedious to prepare when in their ground states, but more easily made using light-induced (photochemical) reactions that proceed through electronically excited states.

Nevertheless, achieving highly enantioselective photochirogenesis is not a trivial matter, because excited molecules are short-lived and highly reactive, and because it is difficult to precisely control the stereochemistry — the geometrical arrangement of groups in a molecule — of products formed from reactions of excited molecules. The control problem has been overcome using a supramolecular approach[2], in which a 'guest' molecule is fixed into a particular position and orientation within a chiral 'host' environment to enable better stereochemical control of the guest's reactions. Hölzl-Hobmeier and co-workers have developed a new take on supramolecular photochirogenesis that they apply to allenes.

Allenes are organic molecules in which one carbon atom (designated C2) forms a double bond to both of its neighbouring carbon atoms (C1 and C3; Fig. 1). These molecules can assume a type of chirality known as axial chirality if two different groups are attached at each of C1 and C3. In their study, Hölzl-Hobmeier and colleagues used axially chiral allenes that have a lactam group attached at C1. The enantiomeric form of these allenes is fixed when they are in their ground states, but they spontaneously interconvert between the two enantiomers when excited to a state known as a triplet[3].

The lactam group is designed to form pairs of hydrogen bonds with a molecule known as an enantiomeric photosensitizing template (T), which was developed previously by workers from the same research group[4]. T forms complexes with the allene, within which it can absorb visible light and transfer the energy to the allene, exciting the latter to a triplet state[5].

So how does T induce the conversion of a racemic mixture of allenes into a single enantiomer, a process known as deracemization? The process begins with the formation of a complex of T with either of the two enantiomers (which are known as the *S*- and *R*-enantiomers, hereafter referred to simply as *S* and *R*; Fig. 2). Irradiation of the resulting complex T·*S* (or T·*R*) by visible light excites T into a triplet state (T*), which then transfers energy to *S* (or *R*). The excited molecule *S** (or *R**) is then released from the complex, regenerating free T for another catalytic cycle. The liberated excited molecules then undergo racemization, and — in the absence of any factors that discriminate between the two enantiomers — eventually relax to form both *S* and *R* products in the ground state in equal quantities.

However, the overall deracemization process can be enantioselective depending on two things: one is how strongly T binds to *S* to form a complex compared with how strongly it binds to *R*; the other is the relative rate at which energy is transferred from T* to *S* and from T* to *R*. For an allene that carries an extremely bulky group known as a *tert*-butyl, Hölzl-Hobmeier and colleagues' experiments show that T binds to *S* about five times more strongly than it does to *R*. This makes sense in the context of the authors' computational simulations, which show that *S* and *R* stack above T in their respective complexes, but that *S* stacks more closely to T than *R* does (see Fig. 4a of the paper[1]) — which suggests that T·*S* is thermodynamically more stable than T·*R*.

Moreover, the enantiomeric excess (e.e.) — a measure of the ratio of enantiomers in a sample of a chiral compound, where 100% indicates the presence of just one enantiomer — reported by Hölzl-Hobmeier *et al.* for the deracemized allene is 96% in favour of *R*. The rate of energy transfer for T* to each of the two enantiomers can therefore be calculated, and it emerges that the rate of energy transfer to *S* is about ten times the rate of transfer from T* to *R*.

The chiral environment generated by T for the allene in the complex therefore has dual, synergistic roles that lead to the extraordinarily high e.e.: when T is in its ground state,

© 2018 Springer Nature Limited. All rights reserved.

there is a thermodynamic preference for it to bind to *S* rather than to *R*; and when T is in its excited state, kinetic factors greatly favour energy transfer to *S* compared with transfer to *R*. Impressively, the authors demonstrated that they could even use their method to convert a sample of the *S*-isomer of the *tert*-butyl-bearing allene (which had an e.e. of 95%) to the *R*-isomer (which had an e.e. of 96%).

Hölzl-Hobmeier and colleagues did not make a wide survey of which chemical groups could be attached to the allenes without disrupting the enantioselectivity of the deracemization, so this remains to be explored. However, groups that could disturb the hydrogen bonding between T and the allene would need to be avoided or protected (temporarily converted into another group that does not interfere with the hydrogen bonding). Nevertheless, the authors show that 17 racemic allenes bearing a variety of groups (see Fig. 3 of the paper[1]) can be deracemized to produce single enantiomers of 89–97% e.e. in good to excellent chemical yields (52–100%). These e.e. values far exceed the value (3.4%) obtained for the first reported deracemization of an allene[6] in the early days of photochirogenesis research. Another attractive feature of the new method is that it uses a small amount of catalyst (only 2.5 mol% compared with the amount of allene used).

The authors' findings unequivocally demonstrate that supramolecular photochirogenesis, when appropriately designed, can be a powerful tool for asymmetric synthesis that cannot be achieved using conventional, heat-activated reactions. The new reactions might be limited by the need to append hydrogen-bonding groups to both the substrate and the photosensitizing template, and by the narrow range of compounds to which they are immediately applicable (which include sulfoxides and binaphthyl compounds). Nevertheless, the general concept and methodology, as well as the mechanistic details revealed by this study, will generate much discussion and open up fresh avenues of research. ■

Cheng Yang *is in the Key Laboratory of Green Chemistry & Technology of Ministry of Education, College of Chemistry, State Key Laboratory of Biotherapy, and Healthy Food Evaluation Research Center, Sichuan University, Chengdu 610064, China.*
Yoshihisa Inoue *is in the Department of Applied Chemistry, Osaka University, Suita 565-0861, Japan.*
e-mail: inoue@chem.eng.osaka-u.ac.jp

1. Hölzl-Hobmeier, A. *et al. Nature* **564**, 240–243 (2018).
2. Yang, C. & Inoue, Y. *Chem. Soc. Rev.* **43**, 4123–4143 (2014).
3. Rodriguez, O. & Morrison, H. *J. Chem. Soc. D Chem. Commun.* 679 (1971).
4. Alonso, R. & Bach, T. *Angew. Chem. Int. Edn* **53**, 4368–4371 (2014).
5. Ley, C., Morlet-Savary, F., Jacques, P. & Fouassier, J. P. *Chem. Phys.* **255**, 335–346 (2000).
6. Drucker, C. S., Toscano, V. G. & Weiss, R. G. *J. Am. Chem. Soc.* **95**, 6482–6484 (1973).

MICROBIOLOGY

Bacterial molecules target viral DNA

Bacteria can use specific protein–based strategies to defend individual cells against viruses. Evidence that bacterial small molecules also target viruses provides fresh insights into how bacteria thwart viral infection. SEE LETTER P.283

MARTHA R. J. CLOKIE

To enjoy beautiful environments, we might need to defend ourselves against the resident pests, from midge flies on Scottish hillsides to mosquitoes in tropical jungles. If pests are numerous and diverse, a broad-spectrum defence strategy, such as spraying an insect repellent, can be best. Bacteria can also use general defences to combat their viral predators, in addition to having a plethora of more-specific defences that target particular viruses. On page 283, Kronheim *et al.*[1] report their analysis of an antiviral defence system that can protect more than one bacterial species. These findings could have major implications for our understanding of how bacteria and viruses interact.

Viruses that infect bacteria are known as bacteriophages, or just phages, and they have key roles in shaping bacterial evolution, population dynamics and physiology. Phages are considered to be the most abundant and diverse biological entities on Earth[2], and it is essential to consider them when trying to gain a full understanding of the bacterial world. Yet despite their importance, there are huge gaps in our knowledge. In many cases, information about phage host ranges (the types of bacterium that a particular phage can infect) is limited. Certain aspects of how bacteria defend themselves against phage attack are also mysterious.

Most bacterial species make numerous and diverse metabolites (small-molecule products of metabolism) that can provide widespread protection against attack from fungi and other types of bacterium. By contrast, most of the well-understood anti-phage defences in bacteria involve proteins, which often offer protection only at the level of the individual cell that makes the protein, rather than providing protection for a bacterial population. One such common bacterial defence is modification of the microbial cell surface to prevent phage attachment. Another strategy, called the CRISPR–Cas defence system[3], depends on an infected bacterium recognizing and capturing sequences from the viral genome and using these to prime a response that kills viruses containing a copy of the captured sequences. Some bacteria take the approach of adding methyl groups to their DNA and degrading all unmethylated, and therefore foreign, DNA[4]. Many other fascinating examples of these 'single-cell' defence strategies exist[5].

Broad-spectrum antiviral defence mechanisms in bacteria do occur but are less well known. For example, bacteria can shed vesicles from their outer membranes to 'mop up' phages[6]. The shortfall of examples in this category probably reflects the limited scope of previous research rather than a lack of such systems per se. Bacteria and phages have coevolved over approximately 3.9 billion years[7], so it seems reasonable to speculate that nonspecific mechanisms might have a key role in bacterial defences. Arguably, such broad-based systems might have a longer evolutionary history than do the more-specific types of defence, and might have shaped the development of the subsequent targeted strategies.

Kronheim and colleagues began to investigate how bacteria might target phages by testing the ability of a total of 4,960 molecules from a drug-discovery library to prevent a phage called lambda from infecting the model bacterium *Escherichia coli*. This revealed 11 molecules that can limit the success of phage infection. Nine of these can embed within DNA and are called DNA-intercalating agents. Out of the 11 molecules, 4 belong to a group known as the anthracyclines. These include the naturally occurring compounds daunorubicin and doxorubicin, which are used as anticancer drugs. The dual ability of these molecules to target cancer cells and phages raises the question of whether they act by recognizing modified DNA.

"One type of molecule can defend diverse species of bacterium against many different types of phage."

The anti-phage effects of daunorubicin and doxorubicin were first discovered more than 50 years ago[8–10]. Yet, strangely, insights[8–10] that bacteria can produce DNA-intercalating agents that target phages did not come to prominence. Research[11–14] from the 1940s and 1950s also demonstrated that several other antibiotics could prevent phage infection. However, these observations were not interpreted as an indication that the molecules were

© 2018 Springer Nature Limited. All rights reserved.

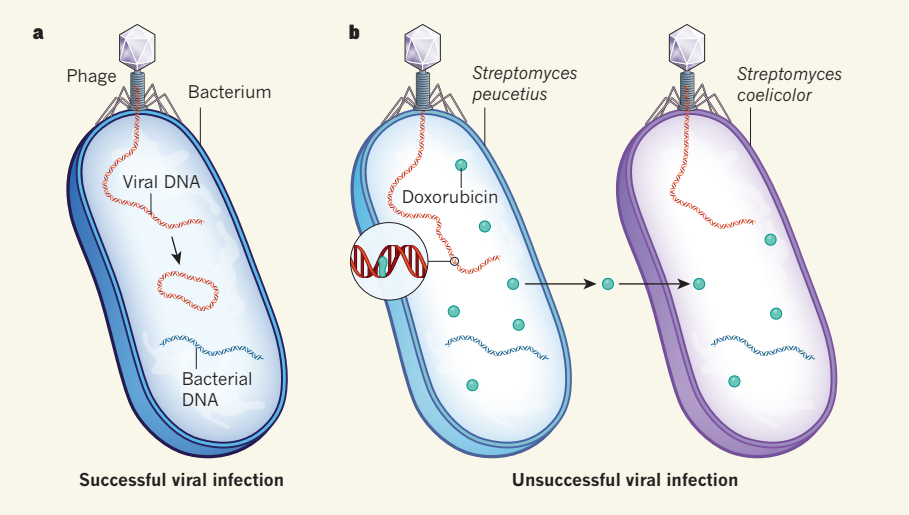


Figure 1 | A bacterial defence approach uses molecules to target viral DNA. **a**, During an early step in the infection of a bacterium by viruses called phages, the linear viral DNA (red) becomes circularized. **b**, Kronheim *et al.*[1] report that bacteria from the genus *Streptomyces* make molecules that can block a successful viral infection. The molecules they identified with this property, such as doxorubicin, are DNA-intercalating compounds — they can become embedded in DNA. These intercalating molecules seem to affect viral rather than bacterial DNA. The authors' results suggest that these molecules block an early stage of viral infection, which might be the step at which viral DNA becomes circularized. Kronheim and colleagues report experimental results consistent with a model in which doxorubicin present in the medium of a culture of *Streptomyces peucetius* bacteria can enter *Streptomyces coelicolor* bacteria and protect them from phage infection. This reveals a broad-spectrum defence mechanism that could offer antiviral protection for bacterial populations from multiple species.

components of a natural bacterial anti-phage defence strategy.

Kronheim *et al.* sought to establish how the molecules they identified act to block phage infection. They demonstrated that viral entry into the cell, viral DNA replication, viral protein synthesis and virus assembly are not inhibited by the addition of daunorubicin. However, they found that daunorubicin can block a step immediately after viral entry and before replication. There will undoubtedly be future studies to determine the mechanism of molecular action at this stage. The most plausible hypothesis suggested by the authors is that daunorubicin blocks the circularization of linear viral DNA. If so, viral DNA that remains in a linear form might be degraded by the host bacterium, or phage infection might be suppressed because the viral DNA cannot interact with the proteins needed for its transcription.

Bacteria from the genus *Streptomyces* are particularly prolific metabolite producers and the source of numerous antibiotics. Kronheim and colleagues provide a crucial demonstration that *Streptomyces* species can produce daunorubicin and doxorubicin, revealing that bacteria can make their own metabolite-based anti-phage system. The authors showed that *Streptomyces* produce many anthracycline-like compounds, some of which prevent the infection of bacteria by specific phages, whereas others prevent infection by a range of phages. The authors tested samples of small-molecule extracts from *Streptomyces* species, and found that 30% of the extracts inhibited phage infection but did not affect bacterial growth, suggesting that bacterial DNA is not susceptible to interference by the molecules that hinder phage infection.

The authors' results suggest that *Streptomyces* bacteria release anthracyclines that can diffuse out of the bacterial cell into the external environment, enter neighbouring bacterial cells, and inhibit phage infection. This was confirmed by adding the medium from several-day-old cultures of *Streptomyces* to fresh cultures of *Streptomyces* to which phages were added. Remarkably, they showed that when doxorubicin-containing, microbe-free medium from cultures of *Streptomyces peucetius* bacteria was added to cultures of *Streptomyces coelicolor* bacteria, it protected *S. coelicolor* from phage infection (Fig. 1).

The defence mechanism uncovered by Kronheim and colleagues contrasts with most known bacterial antiviral mechanisms in two ways. First, it uses metabolites rather than proteins. Second, the mechanism protects not just the cell that produces the anti-phage molecule, but also neighbouring bacterial cells of the same, and even different, bacterial species. This suggests a broad-spectrum metabolite-based defence system that acts in a manner akin to a 'phagicide', whereby one type of molecule can defend diverse species of bacterium against many different types of phage.

By expanding the concept of phage defence from the individual cell to the masses, Kronheim and colleagues' study suggests that phage defence is not as targeted to particular phages as was previously thought. Their work raises many questions. How important is this mechanism, and how are these metabolites made? Are they continually produced, or made only in response to phage infection? It would be interesting to learn how many different types of metabolite are able to target phages, how specific the metabolites' modes of action are, and to what extent such molecules can provide protection across different bacterial species.

Phages have developed ways to overcome most bacterial defences, and can cooperate to evade CRISPR–Cas defences[15,16], so it seems probable that some phages might have developed ways to combat these bacterial defence molecules. Investigating whether this is the case should provide some interesting insights. The concept of phagicides is likely to spark searches for other types of anti-phage metabolite, perhaps leading to the discovery of antiviral metabolites that target other sorts of phage, such as those that have their genetic information in the form of RNA rather than DNA.

Kronheim and colleagues' work also adds to the growing body of evidence revealing the complexity of interactions between phages and bacteria. It follows other paradigm-shifting observations in this research area, such as the report that signalling between phages can affect whether the viruses enter a dormant state or replicate[17]. Building on Kronheim and colleagues' work, it is now time to consider the idea that metabolites can move from bacterium to bacterium to block phage infection. As additional systems are studied, this should help to unravel the extent of this communication, and illuminate how bacteria and their viral predators shape the world in which we live. ■

Martha R. J. Clokie *is in the Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK.*
e-mail: mrjc1@le.ac.uk

1. Kronheim, S. *et al. Nature* **564**, 283–286 (2018).
2. Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. *Bacteriophage* **1**, 31–45 (2011).
3. Makarova, K. S. *et al. Nature Rev. Microbiol.* **12**, 722–736 (2015).
4. Labrie, S. J., Samson, J. E. & Moineau, S. *Nature Rev. Microbiol.* **8**, 317–327 (2010).
5. Doron, S. *et al. Science* **359**, eaar4120 (2108).
6. Manning, A. J. & Kuehn, M. J. *BMC Microbiol.* **1**, 258 (2011).
7. Bamford, D. H. *Res. Microbiol.* **154**, 231–236 (2003).
8. Sanfilippo, A. & Mazzoleni, R. *Giorn. Microbiol.* **12**, 83–92 (1964).
9. Parisi, B. & Soller, A. *Giorn. Microbiol.* **12**, 183–194 (1964).
10. Morita, J., Tanaka, A., Komano, T. & Oki, T. *Agric. Biol. Chem.* **43**, 2629–2631 (1979).
11. Weiss, U., Strelitz, F., Flon, H. & Asheshov, I. N. *Arch. Biochem. Biophys.* **74**, 150–157 (1958).
12. Strelitz, F., Flon, H., Weiss, U. & Asheshov, I. N. *J. Bacteriol.* **72**, 90–94 (1956).
13. Strelitz, F., Flon, H. & Asheshov, I. N. *J. Bacteriol.* **69**, 280–283 (1955).
14. Asheshov, I. N., Strelitz, F., Hall, E. & Flon, H. *Antibiot. Chemother. (Northfield, Ill.)* **4**, 380–394 (1954).
15. Landsberger, M. *et al. Cell* **174**, 908–916 (2018).
16. Borges, A. L. *et al. Cell* **174**, 917–925 (2018).
17. Erez, A. *et al. Nature* **541**, 488–493 (2017).

This article was published online on 5 December 2018.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0776-9

Increased variability of eastern Pacific El Niño under greenhouse warming

Wenju Cai[1,2]*, Guojian Wang[1,2], Boris Dewitte[3,4,5,6], Lixin Wu[1]*, Agus Santoso[2,7], Ken Takahashi[8], Yun Yang[9], Aude Carréric[6] & Michael J. McPhaden[10]

The El Niño–Southern Oscillation (ENSO) is the dominant and most consequential climate variation on Earth, and is characterized by warming of equatorial Pacific sea surface temperatures (SSTs) during the El Niño phase and cooling during the La Niña phase. ENSO events tend to have a centre—corresponding to the location of the maximum SST anomaly—in either the central equatorial Pacific (5° S–5° N, 160° E–150° W) or the eastern equatorial Pacific (5° S–5° N, 150°–90° W); these two distinct types of ENSO event are referred to as the CP-ENSO and EP-ENSO regimes, respectively. How the ENSO may change under future greenhouse warming is unknown, owing to a lack of inter-model agreement over the response of SSTs in the eastern equatorial Pacific to such warming. Here we find a robust increase in future EP-ENSO SST variability among CMIP5 climate models that simulate the two distinct ENSO regimes. We show that the EP-ENSO SST anomaly pattern and its centre differ greatly from one model to another, and therefore cannot be well represented by a single SST 'index' at the observed centre. However, although the locations of the anomaly centres differ in each model, we find a robust increase in SST variability at each anomaly centre across the majority of models considered. This increase in variability is largely due to greenhouse-warming-induced intensification of upper-ocean stratification in the equatorial Pacific, which enhances ocean–atmosphere coupling. An increase in SST variance implies an increase in the number of 'strong' EP-El Niño events (corresponding to large SST anomalies) and associated extreme weather events.

Alternating between El Niño and La Niña events, the ENSO affects extreme weather events, ecosystems and agriculture around the world[1–7]. ENSO events vary greatly[8–15]: the EP-ENSO is associated with strong El Niño events and weak cold SST anomalies, and is characterized by the maximum SST anomaly (the SST anomaly centre) being located in the eastern equatorial Pacific (the 'Niño3' region: 5° S–5° N, 150°–90° W); the CP-ENSO is associated with strong or moderate La Niña events and modest El Niño events, and is characterized by the SST anomaly centre being located in the central equatorial Pacific (5° S–5° N, 160° E–150° W). EP-El Niño events are the strongest and most destructive El Niño events. During such events, SST warming in the Niño3 region leads to flooding in southwest USA, Ecuador and northeast Peru, and to droughts in regions that border the western Pacific[1,4]. In extreme cases, the disruption includes substantial loss of marine life in the eastern Pacific, mass bleaching of corals across the Pacific and beyond[2], and movement of the intertropical convergence zone[7] and of the South Pacific convergence zone towards the equator[5,16], inducing catastrophic floods and droughts across the Pacific region[5,7]. Because of these severe effects, determining how EP-El Niño SST variability responds to greenhouse warming is one of the most important issues in climate science. However, over several model generations, there has been no inter-model consensus on future variability using conventional ENSO indices[17–19].

This lack of consensus is despite inter-model agreement on the change in mean state and modest inter-model agreement on the response of CP-ENSO SST variability and on the change in certain characteristics of ENSO extremes. First, faster warming in the eastern equatorial Pacific than in the surrounding regions and in the equatorial Pacific than in the non-equatorial Pacific[20] facilitates an increased frequency of equatorward shifts of the convergence zones and increased rainfall variability, even if SST variability does not change[7,16,21]. The extreme shifts of the convergence zones occur during El Niño events, particularly during strong ones[5,7,16]. Second, El Niño events with eastward-propagating anomalies increase in frequency as a consequence of weakening Walker circulation[22]. Finally, there has been a focus on the response of CP-ENSO SST variability to greenhouse warming. Although the frequency of CP-El Niño events is projected to increase, the robustness of the increased frequency is debated[9,23–25]. On the other hand, a projected faster warming in the surface layer of the ocean than at depths enhances the role of the relatively cold subsurface water in the central Pacific in generating strong La Niña events, leading to an increased frequency of extreme La Niña events under greenhouse warming[26].

The response of EP-ENSO SST variability to greenhouse warming is even more uncertain, owing to the lack of inter-model consensus[17–19]. Previous examinations of this issue focused on SST variability at a fixed location, typically the Niño3 region[17,18]. This approach assumes that models simulate an EP-ENSO SST anomaly centre that can be represented by the Niño3 SST index, as is the case in observations. Here we show that the longitude of EP-ENSO SST anomaly centres differs greatly from one model to another, particularly when considering models that cannot simulate ENSO diversity. However, we also show that there is a robust increase across models in EP-El Niño SST variability at the anomaly centre of each model.

[1]Key Laboratory of Physical Oceanography, Institute for Advanced Ocean Studies, Ocean University of China and Qingdao National Laboratory for Marine Science and Technology, Qingdao, China. [2]Centre for Southern Hemisphere Oceans Research (CSHOR), CSIRO Oceans and Atmosphere, Hobart, Tasmania, Australia. [3]Centro de Estudios Avanzados en Zonas Áridas, Coquimbo, Chile. [4]Departamento de Biología, Facultad de Ciencias del Mar, Universidad Católica del Norte, Coquimbo, Chile. [5]Millennium Nucleus for Ecology and Sustainable Management of Oceanic Islands, Coquimbo, Chile. [6]Laboratoire d'Etudes en Géophysique et Océanographie Spatiales, Toulouse, France. [7]Australian Research Council Centre of Excellence for Climate Extremes, The University of New South Wales, Sydney, New South Wales, Australia. [8]Servicio Nacional de Meteorología e Hidrología del Perú—SENAMHI, Lima, Peru. [9]College of Global Change and Earth System Science, Beijing Normal University and University Corporation for Polar Research, Beijing, China. [10]NOAA/Pacific Marine Environmental Laboratory, Seattle, WA, USA. *e-mail: wenju.cai@csiro.au; lxwu@ouc.edu.cn

© 2018 Springer Nature Limited. All rights reserved.

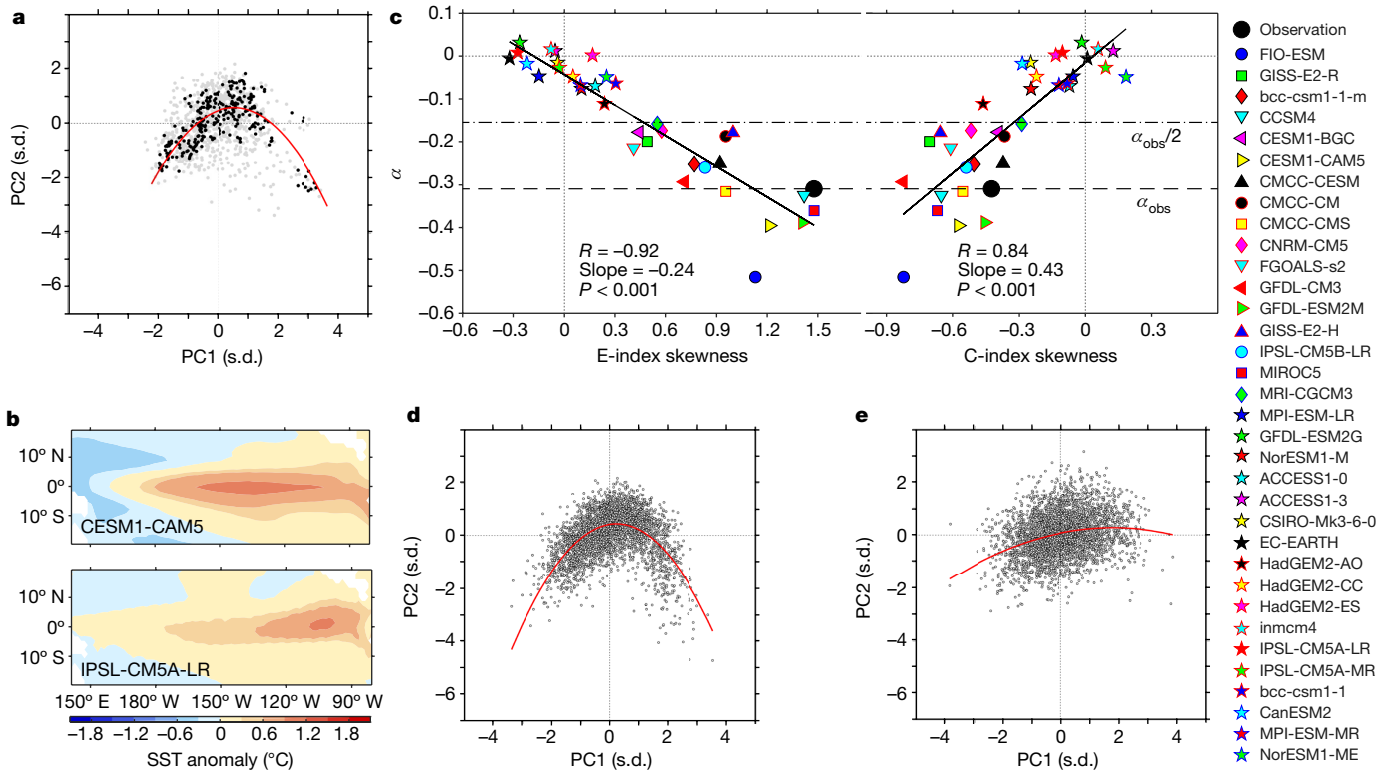


Fig. 1 | Identifying the EP-ENSO anomaly centre in observations and models. a, Nonlinear relationship between the first and second principal components (PC1 and PC2) of SST anomalies averaged over December–February (black dots; see also Extended Data Fig. 1) from five observational reanalysis products. Grey dots indicate monthly data. The nonlinearity is determined by fitting these monthly data with the quadratic function $PC2(t) = \alpha[PC1(t)]^2 + \beta PC1(t) + \gamma$. The red curve shows the same fit, but using the December–February average (black points). **b**, SST anomaly patterns associated with EP-ENSO in two models, highlighting the large difference in the longitude of EP-ENSO anomaly centres (132.25° W for CESM1-CAM5; 101.75° W for IPSL-CM5A-LR) that can occur between climate models. **c**, The parameter α determined using the monthly data versus the skewness of the E-index and C-index for all models analysed (symbols). This parameter is a measure of the contrast between the CP-ENSO and EP-ENSO and of the size of the skewness of the corresponding C-index and E-index. Models with greater $|\alpha|$ systematically produce larger negative skewness in the C-index and larger positive skewness in the E-index. The large black filled circles indicate the observed value α_{obs} (dashed line; the mean of the five observational reanalysis products). The 17 models that produce $|\alpha| < |\alpha_{obs}|/2$ (above the dash-dotted line) are denoted by stars and referred to as 'non-selected'; the other 17 models are shown using various symbols and correspond to the 17 models that we select for further analysis. Details of all models can be found in ref. [29]. The linear fits (solid lines) between α and the E-index or C-index are displayed together with the correlation coefficient R, slope and P value from the regression. **d**, **e**, Nonlinear relationship between the December–February-average principal components for the selected (**d**) and non-selected (**e**) models, with the red curves showing quadratic fits.

Distinguishing SST anomaly centres

At least two ENSO indices are required to distinguish between CP-ENSO and EP-ENSO SST anomaly centres[12,13]. As in previous studies, we use the first two principal modes of an empirical orthogonal function (EOF) analysis of monthly SST anomalies[12,13], with each EOF mode (EOF1 and EOF2) described by a principal spatial pattern, and a principal-component time series scaled to have a variance of unity (see Methods section 'Data, model outputs and EOF analysis'). We applied the EOF analysis to each of five reanalysis products, which we take as the observed SSTs. The positive EOF1 phase exhibits a warm-anomaly centre in the central eastern Pacific; the positive EOF2 phase exhibits a warm-anomaly centre in the central Pacific and a cool-anomaly centre in both the eastern and western parts of the basin[8] (Extended Data Fig. 1a, b). The SST anomaly pattern of an ENSO event is described by a combination of EOF1 and EOF2.

The two monthly principal-component time series display a nonlinear (quadratic[13]) relationship between the two principal modes (Fig. 1a): $PC2(t) = \alpha[PC1(t)]^2 + \beta PC1(t) + \gamma$. For the observations, obtained from multi-reanalysis products (see Methods section 'Data, model outputs and EOF analysis'), the mean value of α is −0.31, which is significantly different from zero at a confidence level of greater than 95%. An EP-ENSO event is described by an E-index[12], which is defined as $(PC1-PC2)/\sqrt{2}$ so that the associated warm-anomaly centre averaged over the season in which an EP-El Niño peaks (December–February) is in the eastern equatorial Pacific. The CP-ENSO regime is described by a C-index, defined as $(PC1+PC2)/\sqrt{2}$, which has a warm-anomaly centre in the central equatorial Pacific (Extended Data Fig. 1c, d). The skewness of the observed monthly E-index is 1.48, reflecting a greater amplitude of EP-El Niño events than of cold SST anomalies. By contrast, the skewness of the monthly C-index is −0.43, reflecting a stronger amplitude of CP-La Niña events than of CP-El Niño events.

Nonlinear dynamics generate skewness

The skewness of the C-index and E-index distributions encapsulates the asymmetry in their associated spatial patterns, and the physical processes responsible for ENSO diversity. Over the central Pacific, during a CP-El Niño event, eastward displacement of the atmospheric convection over the western Pacific warm pool is small, and zonal advection feedback dominates over other processes such as the thermocline feedback[10,11,13,14]. An extreme La Niña event occurs when the central equatorial Pacific thermocline is shallower than normal. This is often associated with the aftermath of an EP-El Niño heat discharge, during which Ekman pumping and nonlinear zonal advection are important, facilitating negative SST skewness[26]. Over the eastern equatorial Pacific, the formation of cool anomalies is curtailed by limited upward displacement of the climatological thermocline, which is already very close to the ocean surface. Instead, this climatological

© 2018 Springer Nature Limited. All rights reserved.

setting of the thermocline favours thermocline deepening and thus the development of strong warm anomalies during EP-El Niño events, which also involve substantial eastward movement of atmospheric deep convection from the western Pacific warm pool to the eastern equatorial Pacific.

These processes in the eastern Pacific region are enhanced by nonlinear Bjerknes feedback, by which the response of zonal winds increases with positive SST anomalies, contributing to the positive SST skewness in the eastern equatorial Pacific[13,14]. We obtained the EP-El Niño and CP-El Niño SST anomaly pattern using a bi-linear regression of the quadratically de-trended December–February-average SST anomaly at each grid point onto the December–February-average E-index and C-index (Extended Data Fig. 1c, d), to focus on the peak ENSO season. The same was carried out for zonal wind stress anomalies (see Methods section 'Diagnosis of nonlinear Bjerknes feedback'), but using monthly data to take into account that winds are important before the peak season. We then took the time series at the associated wind-stress anomaly centre (the longitude of the maximum wind anomalies averaged over 5° S–5° N) to illustrate this nonlinear process. Wind-stress anomalies respond linearly to concurrent monthly SST anomalies in the CP-ENSO centre (Extended Data Fig. 1f). However, the response is nonlinear for EP-ENSO anomalies[13,14,27,28]: stronger for warm anomalies than for cold anomalies (Extended Data Fig. 1e). Enhanced westerly-wind anomalies induce a reduction in equatorial upwelling, an eastward tilting thermocline and westward upper-ocean currents, through Ekman pumping, zonal advection and, particularly, thermocline feedbacks, which promote further growth of eastern Pacific warm anomalies[14,27,28].

Skewness determines the anomaly centre

To identify the EP-ENSO anomaly centre in models, we conducted a similar analysis for 34 CMIP5 models. We applied EOF analysis to monthly SST anomalies, quadratically de-trended over the full period 1900–2099 (see Methods section 'Data, model outputs and EOF analysis'). These models were forced with historical anthropogenic and natural forcings until 2005, and the Intergovernmental Panel on Climate Change Representative Concentration Pathway (RCP) 8.5 future greenhouse gas concentration trajectory from 2006 onwards[29]. Although the multi-model average position of the EP-ENSO anomaly centre compares well with the observed position, the position of the anomaly centre differs from one model to another with a range of 61.5° in longitude (Extended Data Table 1), and the associated SST anomaly patterns could be very different (Fig. 1b). Our approach does not impose a fixed-location anomaly centre, in contrast to approaches that use the Niño3 index. This allows us to assess the response of the EP-ENSO SST simulated by each individual model.

By definition, the model-predicted SST anomaly centres associated with the C-index and E-index correspond to the maximum negative and positive skewness of these indices, respectively. Thus, identifying the CP-ENSO and EP-ENSO anomaly centre in a model is equivalent to locating the maximum negative and positive skewness, assuming that the model is able to generate skewness. An inter-model relationship shows that models with a larger $|\alpha|$ systematically produce greater positive skewness in the E-index (correlation of 0.92) and greater negative skewness in the C-index (correlation of 0.84; Fig. 1c). Thus, the parameter α connects the two skewness values and measures the diversity of the ENSO, which encapsulates the nonlinear Bjerknes feedback discussed above. The parameter α is also related to the response of zonal winds to eastern Pacific warm anomalies and to central Pacific cool anomalies (Extended Data Fig. 2). We therefore select models on the basis of their corresponding value of α.

The majority of models examined here simulate a value of $|\alpha|$ that is lower than the observed value, consistent with previous findings that many models struggle to simulate the two types of ENSO[30,31], and the observed α[32]. Only 17 models produce a value of $|\alpha|$ that is at least 50% of the observed value and generate a reasonable, nonlinear PC1–PC2 relationship (Fig. 1d; see Extended Data Fig. 3 for anomaly patterns and Extended Data Fig. 4 for the nonlinear relationship in some individual models). Excluding these 17 values, the next largest value of $|\alpha|$ is only 36% of the observed value. We therefore use only these 17 models to assess changes in EP-ENSO under greenhouse warming. The remaining 17 models (stars in Fig. 1c) generate small values of $|\alpha|$—with some values of α having the opposite sign in at least one centre compared to the observed value (Fig. 1c)—and consistently produce a far weaker nonlinear PC1–PC2 relationship (Fig. 1e), indicating a weaker or lack of nonlinear Bjerknes feedback (see Extended Data Fig. 5 for anomaly patterns and Extended Data Fig. 6 for the lack of a nonlinear relationship in some non-selected individual models). In these non-selected models, PC1 and PC2 are scattered without a well-defined relationship, which means that events with the same E-index can correspond to a combination of large PC1 and small PC2, or the other way around. As a result, the EP-El Niño anomaly pattern and the location of the anomaly centre vary substantially from one event to another, so it is difficult to assess the future change in variability.

Variability increases at the EP–ENSO centre

We compare the standard deviation (s.d.) of the E-index in the present-day control (1900–1999) and future climate change (2000–2099) periods, each of 100 years. 15 of the 17 selected models (88%) simulate an increased variance in the E-index in the future period (red bars Fig. 2a). The two models that generate reduced EP-El Niño variability (CCSM4[33] and GFDL-ESM2M) also produce reduced climatological rainfall in the equatorial eastern Pacific, in contrast to increased climatological rainfall in the ensemble average. However, it is not clear whether the reduced climatological rainfall is a consequence or a cause of the decreased EP-El Niño variability. The ensemble-mean increase in the standard deviation of EP-El Niño SST is 15%, which is significant at more than the 95% confidence level according to a bootstrap test (see Methods section 'Statistical significance test'; Extended Data Fig. 7a). The increase in variance translates to a 25% and 27% increase in occurrences of EP-El Niño events with an E-index of more than 0.75 s.d. and more than 1 s.d., respectively. For strong events (E-index $>$ 1.5 s.d.; Fig. 2b), the increase in frequency is 47%, although there is no inter-model consensus on changes in intensity. By contrast, for the non-selected models, there is no inter-model consensus on the change in variance, with only 9 of the 17 models (53%) producing an increase (Extended Data Fig. 8a).

Sensitivity to emission scenarios and to model generations suggests that 12 of the 15 selected CMIP5 models (80%) that are forced under RCP4.5 and five of the seven CMIP3 models (71%) forced under the A2 scenario and selected using the same value of α generate an increase in E-index variance (see Methods section 'Sensitivity to emission scenarios'). In addition, a sensitivity test of our finding to model selection reveals that, even when all 34 CMIP5 models under RCP8.5 are considered, there is still a reasonable inter-model consensus on the increased E-index variance, with 24 of the 34 models (71%) simulating an increase (Extended Data Fig. 8a). By contrast, in terms of Niño3 SST variance, only 18 of the 34 CMIP5 models (53%) produce an increase (Extended Data Fig. 8b). In the selected models, the inter-model consensus is further enhanced because the CP-ENSO and EP-ENSO regimes are more distinguishable, with the EP-ENSO anomaly pattern and centre better defined than in the non-selected models, as indicated by the large $|\alpha|$ and SST skewness values. As such, for the selected models, the longitudinal range of the EP-ENSO anomaly centre is reduced from 61.5° to 28.5° and the Niño3 SST anomaly becomes a reasonable index to represent EP-El Niño events. 12 of the 17 selected models (70.5%) produce an increased Niño3 SST variance. This is a reasonable inter-model consensus, although still lower than that using the E-index (Extended Data Fig. 8a, b).

Stratification change boosts dynamic coupling

We find that although there is an ensemble-mean increase in $|\alpha|$—which signifies the continuous separation of eastern and central Pacific

© 2018 Springer Nature Limited. All rights reserved.

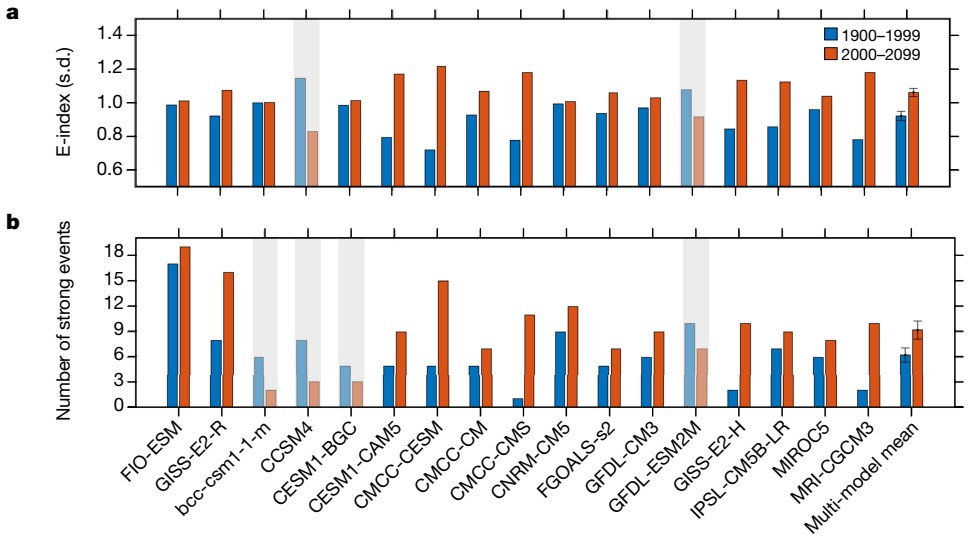


Fig. 2 | Projected increase in EP-ENSO variance. a, Comparison of the standard deviation of the E-index over the present-day (1900–1999) and future (2000–2099) 100-year periods in the 17 selected models. 15 of the 17 selected models (88%) simulate a greater variance in the E-index in the future period (red bars) than in the present-day period (blue bars); the two models that simulate a reduction in variance are greyed out. **b**, Number of strong EP-El Niño events (E-index > 1.5 s.d.) that occurred in the two 100-year periods. The multi-model mean is also shown in **a** and **b**; error bars in the multi-model mean correspond to the 95% confidence interval. The differences between the present-day and future multi-model-mean E-index (s.d.) and between the present-day and future multi-model-mean number of strong events are statistically significant at more than the 95% confidence level. The increase in EP-ENSO SST variance (E-index variance) generally translates to more EP-El Niño events for a given E-index intensity.

centres—from the present-day to the future climate, the change is statistically insignificant and without an inter-model consensus. On the other hand, the change in mean climate is robust and can explain the increased EP-ENSO variance. For example, under greenhouse warming, the change in mean state includes faster warming in the eastern equatorial Pacific than in the surrounding regions[20] (Extended Data Fig. 9a). Although its direct effect on SST variability is already removed through the quadratic de-trending process, there is a statistically significant relationship between the intensity of the warming pattern and eastern Pacific variability (Extended Data Fig. 9b). The surface warming pattern, with stronger warming in the eastern equatorial Pacific than in the surrounding regions, contributes to the increased EP-El Niño variability by facilitating more frequent atmospheric convection in the region.

However, a greater contribution to this variability comes from increased vertical stratification of the upper equatorial Pacific Ocean (Fig. 3a, b). The increased vertical stratification is another robust feature of the change in mean state that is supported by a strong inter-model consensus[17,18]. To assess the effect of the increased stratification, we conduct a vertical mode decomposition of the mean Brunt Väisälä frequency profiles (see Methods section 'Wind projection coefficient') and determine the projection of the wind-stress forcing momentum onto the dominant ocean baroclinic modes[34,35] (the wind-projection coefficient), which measures the dynamical coupling between the atmosphere and the ocean at the wind anomaly centre[36–38]. The centre, determined by a bi-linear regression of quadratically de-trended monthly zonal wind anomalies onto the C-index and E-index, is located west of the SST anomaly centre (Extended Data Table 1). Because ENSO instability increases with wind–ocean coupling, stochastic forcing is more likely to trigger positive feedbacks for an El Niño event[38–42]. In all selected models, the coupling increases in the future climate (Fig. 3c). There is a strong inter-model consensus, and models with greater strengthening in vertical stratification at the wind anomaly centre systematically produce a greater increase in the coupling (Fig. 3d). Thus, the increased stratification enhances the EP-El Niño by increasing the dynamical coupling between the ocean and the atmosphere.

Additional analysis reveals that the dynamical coupling at the wind centre of the C-index increases from the present-day to the future climate by a similar amount, suggesting that the same mechanism operates for the CP-ENSO. This is indeed the case (see Methods section 'Response of central Pacific ENSO'). In particular, 11 of the 17 selected models (65%) generate an increased frequency of CP-El Niño, defined as when the magnitude of the C-index is greater than 1 s.d. The inter-model consensus is not as strong as for the EP-El Niño, perhaps in part because there is no faster warming in the central Pacific region to facilitate atmospheric convection and thus enhance SST variability, as there is for the EP-El Niño.

Summary

Our finding of a greenhouse-warming-induced increase in EP-El Niño SST variance is in contrast to previous findings of no consensus using the Niño3 SST index. Previous studies assumed that all models produce an anomaly pattern and centre that can be represented by the Niño3 index, as generally seen in observations. We show that the EP-ENSO pattern and its anomaly centre differ greatly from one model to another, and therefore cannot be represented by the spatially fixed Niño3 SST index. Further, the EP-El Niño SST anomaly centre is determined by the positive-skewness centre, which is governed by the associated nonlinear processes. Focusing on the different EP-El Niño anomaly centres for each model, there is an increase in EP-El Niño SST variance under greenhouse warming, with a strong inter-model consensus. The robust result arises from the use of process-based metrics representing the nonlinear Bjerknes feedback that underlies ENSO diversity. The increased SST variance stems from enhanced stratification of the upper equatorial Pacific Ocean under greenhouse warming, which enhances the wind–ocean coupling that is conducive to an increase in SST anomalies. With this projected increase, we should expect more extreme weather events associated with the EP-El Niño, with important implications for twenty-first-century climate, extreme weather and ecosystems.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0776-9.

© 2018 Springer Nature Limited. All rights reserved.

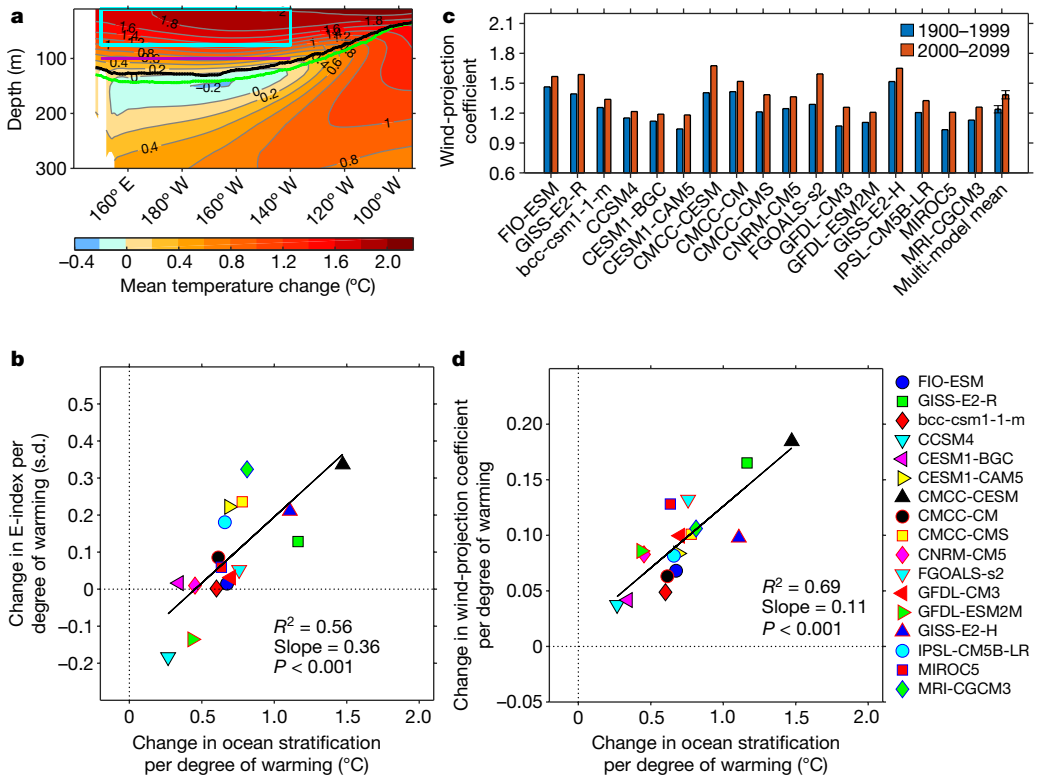


Fig. 3 | Mechanism for the projected increase in EP-ENSO variance. **a**, Multi-model-mean change in equatorial ocean temperature (the upper 300 m) between future (2000–2099) and present-day (1900–1999) climates (colour scale; values are also indicated on each contour). The present-day (green) and future (black) thermoclines are also shown. The stratification increases and the thermocline shallows under greenhouse warming. **b**, Statistically significant ($P < 0.001$) relationship between the change (between future and present-day climates) in ocean stratification and the change (between future and present-day climates) in E-index. The ocean stratification is calculated as the difference between the mean temperature over the upper 75 m (cyan box in **a**) and the temperature at 100 m (purple line in **a**), both averaged over the longitude range 150° E–140° W. To enhance the inter-model comparability, the changes are scaled by the increase in global-mean temperature over the present-day and future periods. Approximately 55% of the change in E-index variance is attributable to the change in ocean stratification, much more than is due to the faster surface warming in the eastern equatorial Pacific (Extended Data Fig. 9). **c**, Comparison of the wind–ocean coupling coefficients (wind-projection coefficients) in present-day and future climates. All models generate increased coupling in the future. ENSO instability increases with wind–ocean coupling. **d**, Statistically significant ($P < 0.001$) relationship between greenhouse-warming-induced changes in ocean stratification and in the wind–ocean coupling coefficient.

Received: 6 August; Accepted: 25 October 2018;
Published online 12 December 2018.

1. Ropelewski, C. F. & Halpert, M. S. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Weath. Rev.* **115**, 1606–1626 (1987).
2. Glynn, P. W. & DE Weerdt, W. H. Elimination of two reef-building hydrocorals following the 1982-83 El Niño warming event. *Science* **253**, 69–71 (1991).
3. Bove, M. C., O'Brien, J. J., Eisner, J. B., Landsea, C. W. & Niu, X. Effect of El Niño on US landfalling hurricanes, revisited. *Bull. Am. Meteorol. Soc.* **79**, 2477–2482 (1998).
4. McPhaden, M. J., Zebiak, S. E. & Glantz, M. H. ENSO as an integrating concept in earth science. *Science* **314**, 1740–1745 (2006).
5. Vincent, E. M. et al. Interannual variability of the South Pacific Convergence Zone and implications for tropical cyclone genesis. *Clim. Dyn.* **36**, 1881–1896 (2011).
6. Valle, C. A. et al. The impact of the 1982–1983 El Niño-Southern Oscillation on seabirds in the Galapagos Islands, Ecuador. *J. Geophys. Res.* **92**, 14437–14444 (1987).
7. Cai, W. et al. Increasing frequency of extreme El Niño events due to greenhouse warming. *Nat. Clim. Chang.* **4**, 111–116 (2014).
8. Ashok, K., Behera, S. K., Rao, S. A., Weng, H. & Yamagata, T. El Niño Modoki and its possible teleconnection. *J. Geophys. Res.* **112**, C11007 (2007).
9. Yeh, S.-W. et al. El Niño in a changing climate. *Nature* **461**, 511–514 (2009).
10. Kug, J.-S., Jin, F.-F. & An, S.-I. Two types of El Niño events: cold tongue El Niño and warm pool El Niño. *J. Clim.* **22**, 1499–1515 (2009).
11. Kao, H. Y. & Yu, J.-Y. Contrasting eastern-Pacific and central-Pacific types of ENSO. *J. Clim.* **22**, 615–632 (2009).
12. Takahashi, K., Montecinos, A., Goubanova, K. & Dewitte, B. ENSO regimes: reinterpreting the canonical and Modoki El Niño. *Geophys. Res. Lett.* **38**, L10704 (2011).
13. Dommenget, D., Bayr, T. & Frauen, C. Analysis of the non-linearity in the pattern and time evolution of El Niño Southern Oscillation. *Clim. Dyn.* **40**, 2825–2847 (2013).
14. Takahashi, K. & Dewitte, B. Strong and moderate nonlinear El Niño regimes. *Clim. Dyn.* **46**, 1627–1645 (2016).
15. Capotondi, A. et al. Understanding ENSO diversity. *Bull. Am. Meteorol. Soc.* **96**, 921–938 (2015).
16. Cai, W. et al. More extreme swings of the South Pacific convergence zone due to greenhouse warming. *Nature* **488**, 365–369 (2012).
17. Cai, W. et al. ENSO and Greenhouse warming. *Nat. Clim. Chang.* **5**, 849–859 (2015).
18. Collins, M. et al. The impact of global warming on the tropical Pacific Ocean and El Niño. *Nat. Geosci.* **3**, 391–397 (2010).
19. Watanabe, M. et al. Uncertainty in the ENSO amplitude change from the past to the future. *Geophys. Res. Lett.* **39**, L20703 (2012).
20. Xie, S.-P. et al. Global warming pattern formation: sea surface temperature and rainfall. *J. Clim.* **23**, 966–986 (2010).
21. Power, S., Delage, F., Chung, C., Kociuba, G. & Keay, K. Robust twenty-first-century projections of El Niño and related precipitation variability. *Nature* **502**, 541–545 (2013).
22. Santoso, A. et al. Late-twentieth-century emergence of the El Niño propagation asymmetry and future projections. *Nature* **504**, 126–130 (2013).
23. McPhaden, M. J., Lee, T. & McClurg, D. El Niño and its relationship to changing background conditions in the tropical Pacific. *Geophys. Res. Lett.* **38**, L15709 (2011).
24. Newman, M., Shin, S.-I. & Alexander, M. A. Natural variation in ENSO flavors. *Geophys. Res. Lett.* **38**, L14705 (2011).
25. Yeh, S.-W., Kirtman, B. P., Kug, J.-S., Park, W. & Latif, M. Natural variability of the central Pacific El Niño event on multi-centennial timescales. *Geophys. Res. Lett.* **38**, L02704 (2011).
26. Cai, W. et al. More frequent extreme La Niña events under greenhouse warming. *Nat. Clim. Chang.* **5**, 132–137 (2015).
27. Choi, K.-Y., Vecchi, G. A. & Wittenberg, A. T. ENSO transition, duration, and amplitude asymmetries: role of the nonlinear wind stress coupling in a conceptual model. *J. Clim.* **26**, 9462–9476 (2013).
28. Frauen, C. & Dommenget, D. El Niño and La Niña amplitude asymmetry caused by atmospheric feedbacks. *Geophys. Res. Lett.* **37**, L18801 (2010).

© 2018 Springer Nature Limited. All rights reserved.

29. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
30. Kim, S. T. & Yu, J.-Y. The two types of ENSO in CMIP5 models. *Geophys. Res. Lett.* **39**, L11704 (2012).
31. Ham, Y.-G. & Kug, J.-S. How well do current climate models simulate two types of El Nino? *Clim. Dyn.* **39**, 383–398 (2012).
32. Karamperidou, C., Jin, F.-F. & Conroy, J. L. The importance of ENSO nonlinearities in tropical Pacific response to external forcing. *Clim. Dyn.* **49**, 2695–2704 (2017).
33. Stevenson, S. et al. Will there be a significant change to El Niño in the twenty-first century? *J. Clim.* **25**, 2129–2145 (2012).
34. Cane, M. A. & Sarachik, E. S. The response of a linear baroclinic equatorial ocean to periodic forcing. *J. Mar. Res.* **39**, 651–693 (1981).
35. Zebiak, S. E. & Cane, M. A. A model El Niño–Southern Oscillation. *Mon. Weath. Rev.* **115**, 2262–2278 (1987).
36. Dewitte, B., Reverdin, G. & Maes, C. Vertical structure of an OGCM simulation of the equatorial Pacific Ocean in 1985–1994. *J. Phys. Oceanogr.* **29**, 1542–1570 (1999).
37. Dewitte, B. et al. Low frequency variability of temperature in the vicinity of the equatorial thermocline in SODA: role of equatorial wave dynamics and ENSO asymmetry. *J. Clim.* **22**, 5783–5795 (2009).
38. An, S.-I. & Jin, F.-F. Collective role of thermocline and zonal advective feedbacks in the ENSO mode. *J. Clim.* **14**, 3421–3432 (2001).
39. Thual, S., Dewitte, B., An, S.-I. & Ayoub, N. Sensitivity of ENSO to stratification in a recharge-discharge conceptual model. *J. Clim.* **4**, 4331–4348 (2011).
40. Thual, S., Dewitte, B., An, S.-I., Illig, S. & Ayoub, N. Influence of recent stratification changes on ENSO stability in a conceptual model of the equatorial Pacific. *J. Clim.* **26**, 4790–4802 (2013).
41. Gent, P. R. & Luyten, J. R. How much energy propagates vertically in the equatorial oceans? *J. Phys. Oceanogr.* **15**, 997–1007 (1985).
42. Kim, W. & Cai, W. The importance of the eastward zonal current for generating extreme El Niño. *Clim. Dyn.* **42**, 3005–3014 (2014).

Acknowledgements This work is supported by the Centre for Southern Hemisphere Oceans Research, a joint research centre between QNLM and CSIRO. W.C., G.W. and A.S. are also supported by the Earth Systems and Climate Change Hub of the Australian Government's National Environmental Science Program, and a CSIRO Office of Chief Executive Science Leader award. B.D. was supported by Fondecyt (grant number 1171861) and LEFE-GMMC. PMEL contribution number: 4817.

Reviewer information *Nature* thanks Y.-G. Ham and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.C. conceived the study and wrote the initial manuscript in collaboration with L.W. and B.D. G.W. performed the model analysis and generated the final figures. B.D. analyses the dynamical coupling between the atmosphere and the ocean. A.S., B.D., K.T., A.C., Y.Y. and M.J.M. contributed to interpretation of the results, discussion of the associated dynamics and improvement of the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0776-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to W.C. or L.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data, model outputs and EOF analysis. We use five SST reanalysis products[43–47] to characterize ENSO diversity, and atmospheric circulation fields from the National Center for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) global reanalysis[48]. The five reanalysis products are: HadISST v1.1 (Hadley Centre Sea Ice and Sea Surface Temperature dataset version 1.1)[43] from 1948 to 2015; ERSST v5 (Extended Reconstructed Sea Surface Temperature version 5)[44] from 1948 to 2015; OISST v2 (NOAA Optimum Interpolation Sea Surface Temperature version 2)[45] from 1982 to 2015; ORA-s3 (ECMWF Ocean Analysis System: ORA-s3)[46] from 1959 to 2009; and ORA-s4 (ECMWF Ocean Analysis System: ORA-s4)[47] from 1958 to 2013. We use a multivariate signal-processing method referred to as EOF analysis[49] in an equatorial domain (15° S–15° N, 140° E–80° W) to de-convolve spatio-temporal variability into orthogonal modes, each described by a principal spatial pattern and an associated principal component (PC) time series. The PC time series is scaled to have a standard deviation of one. For the observational reanalysis products, EOF analysis is applied to monthly SST anomalies, referenced to their long-term mean. The CP-ENSO and EP-ENSO regimes were reconstructed using EOF1 and EOF2 such that their temporal variability can be described by a C-index ($(PC1 + PC2)/\sqrt{2}$) and an E-index ($(PC1-PC2)/\sqrt{2}$), respectively. Each regime is associated with a suite of distinct processes that lead to the negative and positive skewness in the C-index and E-index, respectively, as discussed in the main text. This approach was applied to 34 CMIP5 coupled global climate models (CGCMs; Extended Data Table 1) forced with historical anthropogenic and natural forcings, and future greenhouse gases under the RCP8.5 scenario[29], covering the 200-year period 1900–2099. Monthly anomalies referenced to the climatology of the first 100 years were constructed and quadratically de-trended.

Diagnosis of nonlinear Bjerknes feedback. For each model, we obtained the associated zonal wind-stress anomaly pattern through the same bi-linear regression onto the monthly E-index and C-index to identify the location of the maximum anomaly associated with each index. We used monthly anomalies to obtain the associated wind-stress anomalies (to capture that the wind-stress response is important during the development phase), but for the SST anomaly pattern and centre we focused on the mature phase of December–February. The westerly anomaly centre is located to the west of the SST anomaly centre, consistent with the fact that the warm anomalies are a dynamic consequence of wind-induced eastward-propagating equatorial downwelling Kelvin waves (Extended Data Table 1). Quadratically de-trended monthly wind anomalies at the centre were plotted against the monthly C-index and E-index, using all samples from observations, the 17 selected models and the 17 non-selected models. Samples were binned at a C- or E-index interval of 0.25 s.d. to obtain median values for each bin (Extended Data Figs. 1e, f, 3e, f and 5e, f).

The nonlinear Bjerknes feedback is measured by a ratio of the regression slope for binned values with a positive C-index, or E-index, over the slope for a negative index. A greater ratio indicates strong nonlinear Bjerknes feedback. For the non-selected models, the wind response to SST anomalies associated with the C-index is essentially linear (ratio of 1.10; Extended Data Fig. 5f) and the response to SST anomalies associated with the E-index becomes moderately nonlinear (ratio of 1.69; Extended Data Fig. 5e). For the selected models, the corresponding ratios are 1.09 and 2.49 (Extended Data Fig. 3e, f); that is, there is a much stronger nonlinear response for the EP-ENSO regime. The response to SST anomalies associated with the E-index for the selected models is close to the observed value (2.42; Extended Data Fig. 1e). In other words, in the selected models the CP and EP regimes are far more distinguishable, and EP- and CP-ENSO anomaly centres are more clearly separated and better defined than in the non-selected models. This is reflected in the lack of, or weak, nonlinear relationship between PC1 and PC2 in the non-selected models, in which an EP-El Niño event is a combination of EOF1 and EOF2, which are uncorrelated, such that in a given model there is a strong inter-event diversity in the spatial pattern and the anomaly centre of the EP-El Niño.

Statistical significance test. We use a bootstrap method to examine whether the increased E-index variance is statistically significant. The 17 standard deviation values of the E-index in the present-day period from the 17 selected models are re-sampled randomly to construct 10,000 realizations of mean standard deviation over 17 models. In this random re-sampling process, any model is allowed to be selected again. The standard deviation of the 10,000 inter-realizations of mean standard deviation for the control period is 0.027. The same is carried out for the future period, and the standard deviation of the inter-realization is 0.024. The increased standard deviation in the future period is greater than the sum of these two standard deviation values, indicating statistical significance above the 95% confidence level (Extended Data Fig. 7a). Identical analyses for increased occurrences in EP-El Niño events with E-index > 1.5 s.d. and for increased wind-projection coefficients leads to the same conclusion (Extended Data Fig. 7b, c).

Sensitivity to emission scenarios. 32 CMIP5 models were forced under the RCP4.5 emission scenario and 15 were selected. 12 of the 15 selected models (80%) produced increased E-index variance. Using all models under this scenario, 20 of the 32 models (63%) generate an increase. The same conclusion is found when applying the same analysis and model-selection criterion on an ensemble of 16 CMIP3 models forced under the A2 emission scenario. Seven CMIP3 models were selected and five (71%) produce increased E-index variance.

Wind projection coefficient. From linear theory, the total amount of momentum flux associated with equatorial wave dynamics can be estimated by the zonal wind stress along the equator multiplied by a coefficient, referred to as the wind-projection coefficient P_n, which depends on the vertical stratification of the ocean[50]. This coefficient corresponds to the coupling efficiency between the ocean and the atmosphere associated with equatorial wave dynamics for a particular baroclinic mode n[34,35]. In a multi-mode context, the coefficients associated with the first three baroclinic modes allow us to characterize the mean thermocline shape (sharpness), depth and intensity[39–41]. These coefficients have been used to diagnose the long-term variability in vertical stratification along the equator associated with changes in ENSO amplitude in reanalysis products and CGCMs[36,51,52]. We calculate the wind-projection coefficients from climatological temperature and salinity profiles of the present-day and future climates[36], and calculate the following quantity for both periods for each of the models[53]:

$$P = \sum_{n=1}^{n=3} P_n = \sum_{n=1}^{n=3} \frac{150}{\int_{z=5{,}000\ \mathrm{m}}^{z=0} F_n^2(x_E, z)\mathrm{d}z}$$

Here F_n corresponds to the vertical mode structure and x_E is the location along the equator, where the salinity and temperature profile are considered. This location is taken as the centre of action of the zonal winds stress for the EP regime and corresponds to the maximum amplitude of the regressed patterns of the zonal wind stress associated with the EP regime (Extended Data Table 1). The factor of 150 is a normalizing coefficient (in metres), corresponding to the average thermocline depth in the equatorial CP. The larger the value of P, the sharper the mean thermocline and the larger the input of momentum flux into the baroclinic ocean response.

Warming pattern. CGCMs produce a warming pattern with faster warming in the equatorial EP than in the surrounding regions, with a strong inter-model consensus[17,18,20]. We calculated the warming in each model as the difference between the average over the future and present-day 100-year periods, and normalized it by the difference in the global-mean temperature (in units of warming per degree of global warming). We constructed a multi-model mean over the 17 selected models (Extended Data Fig. 9a) and then projected individual-model warming patterns onto the multi-model-mean warming pattern to obtain inter-model variations in the warming pattern. We examined the relationship between the inter-model warming pattern and change in the E-index (Extended Data Fig. 9b).

Response of central Pacific ENSO. Inter-model consensus on changes in monthly C-index variance is weaker compared to that of the monthly E-index, with 10 of the 17 selected models (59%) producing an increase. Without model selection, 19 of the 34 models (56%) produce an increase.

In terms of changes in the frequency of CP-El Niño events, defined as when the December–February-average C-index is greater than 1 s.d., 11 of the 17 selected models (65%) produce an increase. Without model selection, 24 of the 34 models (71%) produce an increase.

In terms of changes in the frequency of extreme La Niña events, defined as when the magnitude of the December–February-average C-index is greater than 1.75 s.d., 11 of the 17 selected models (65%) produce an increase. Without model selection, 24 of the 34 models (71%) produce an increase. Note that some of the 24 models that contribute to the increased frequency of extreme La Niña events are different from those that contribute to the increased frequency of CP-El Niño events.

Code availability. Codes used to calculate the EOF and α can be downloaded from https://drive.google.com/open?id=1d2R8wKpFNW-vMIfoJsbqIGPIBd9Z_8rj; codes for calculating the wind-projection coefficients using ocean salinity and temperature are available on request.

Data availability

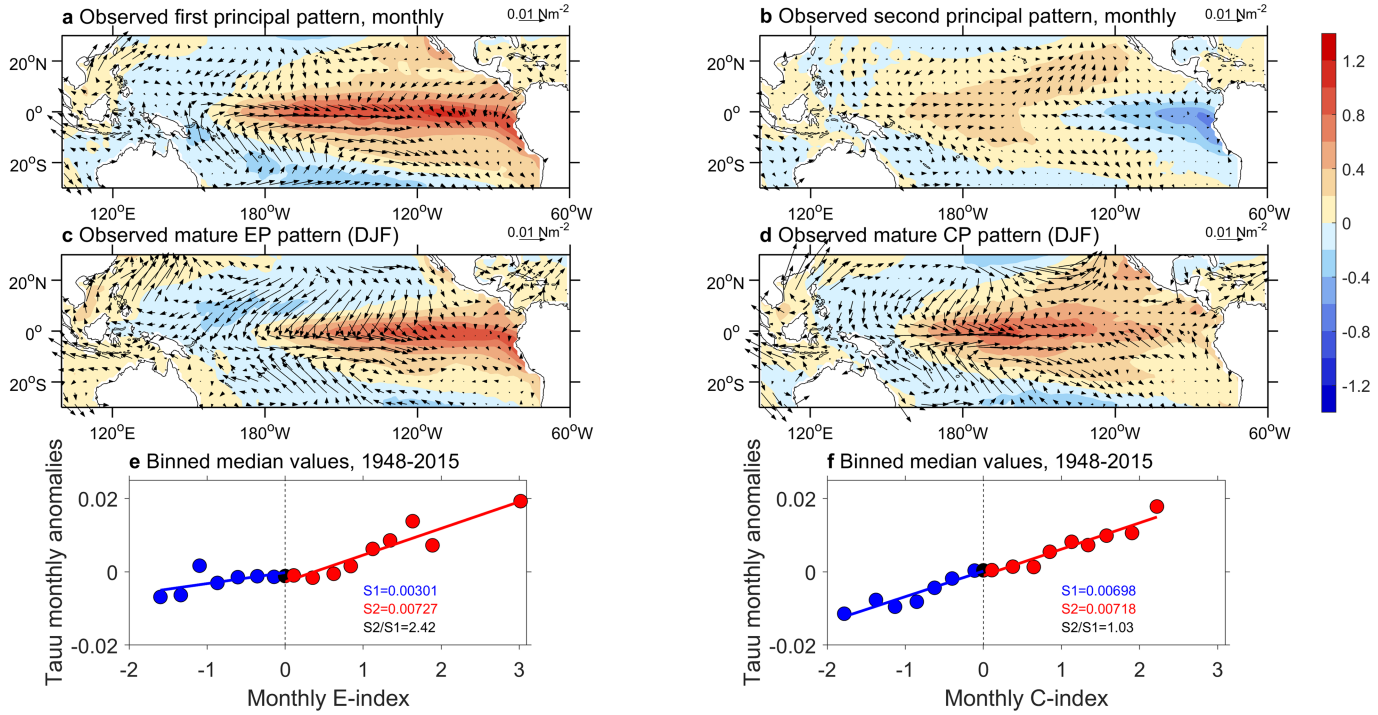
Data related to this paper can be downloaded from the following: HadISST v1.1, https://www.esrl.noaa.gov/psd/data/gridded/data.hadsst.html; ERSST v5, https://www.ncdc.noaa.gov/data-access/marineocean-data/extended-reconstructed-sea-surface-temperature-ersst-v5; OISST v2, https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html; ORA-s3, http://apdrc.soest.hawaii.edu/datadoc/ecmwf_oras3.php; ORA-s4, https://climatedataguide.ucar.edu/climate-data/oras4-ecmwf-ocean-reanalysis-and-derived-ocean-heat-content; and CMIP5 database, http://www.ipcc-data.org/sim/gcm_monthly/AR5/.

43. Rayner, N. A. et al. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**, 4407 (2003).

© 2018 Springer Nature Limited. All rights reserved.

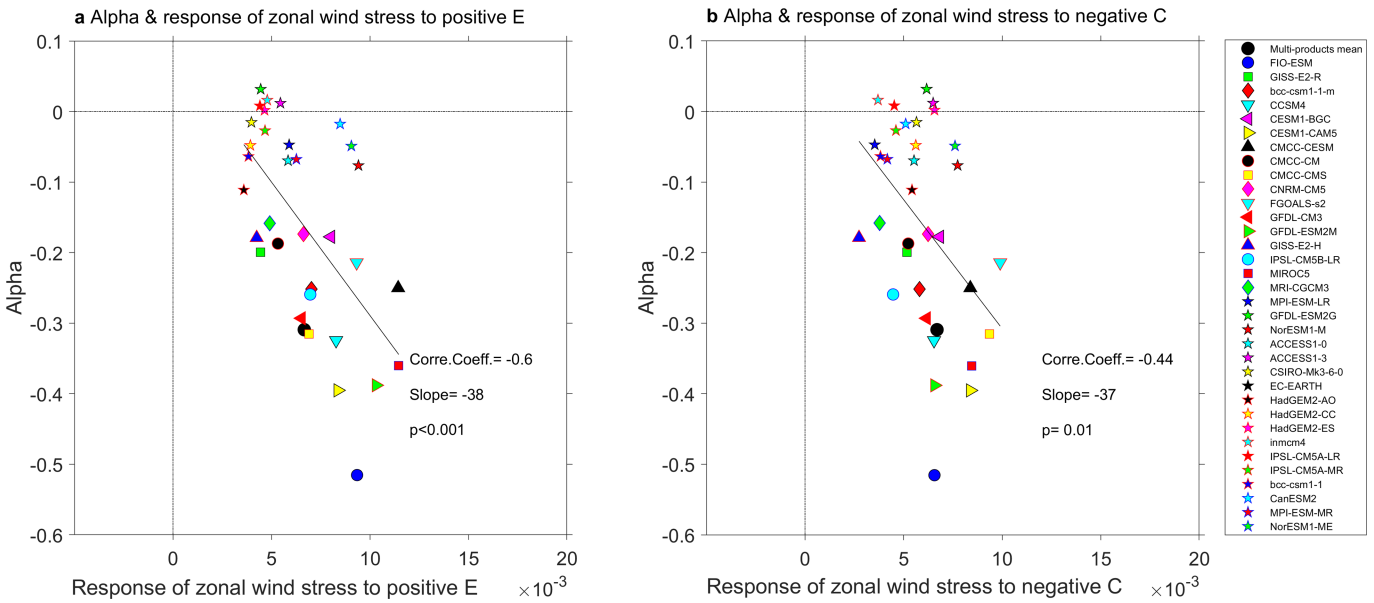
44. Huang, B. et al. Extended reconstructed sea surface temperature version 5 (ERSSTv5), upgrades, validations, and intercomparisons. *J. Clim*. **30**, 8179–8205 (2017).
45. Reynolds, R. W. et al. An improved in situ and satellite SST analysis for climate. *J. Clim*. **15**, 1609–1625 (2002).
46. Balmaseda, M. A., Vidard, A. & Anderson, D. L. T. The ECMWF ocean analysis system: ORA-S3. *Mon. Weath. Rev*. **136**, 3018–3034 (2008).
47. Balmaseda, M. A., Mogensen, K. & Weaver, A. T. Evaluation of the ECMWF ocean reanalysis system ORAS4. *Q. J. R. Meteorol. Soc*. **139**, 1132–1161 (2013).
48. Kalnay, E. et al. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc*. **77**, 437–472 (1996).
49. Lorenz, E. N. *Empirical Orthogonal Functions and Statistical Weather Prediction*. Statistical Forecast Project Report 1 (MIT Department of Meteorology, 1956).
50. Fjelstad, J. E. *Interne Wellen* (Cammermeyer in Komm., 1933).
51. Dewitte, B., Yeh, S.-W., Moon, B.-K., Cibot, C. & Terray, L. Rectification of the ENSO variability by interdecadal changes in the equatorial background mean state in a CGCM simulation. *J. Clim*. **20**, 2002–2021 (2007).
52. Yeh, S.-W., Dewitte, B., Yim, B. Y. & Noh, Y. Role of the upper ocean structure in the response of ENSO-like SST variability to global warming. *Clim. Dyn*. **35**, 355–369 (2010).
53. Blumenthal, M. B. & Cane, M. A. Accounting for parameter uncertainties in model verification: an illustration with tropical sea surface temperature. *J. Phys. Oceanogr*. **19**, 815–830 (1989).

© 2018 Springer Nature Limited. All rights reserved.



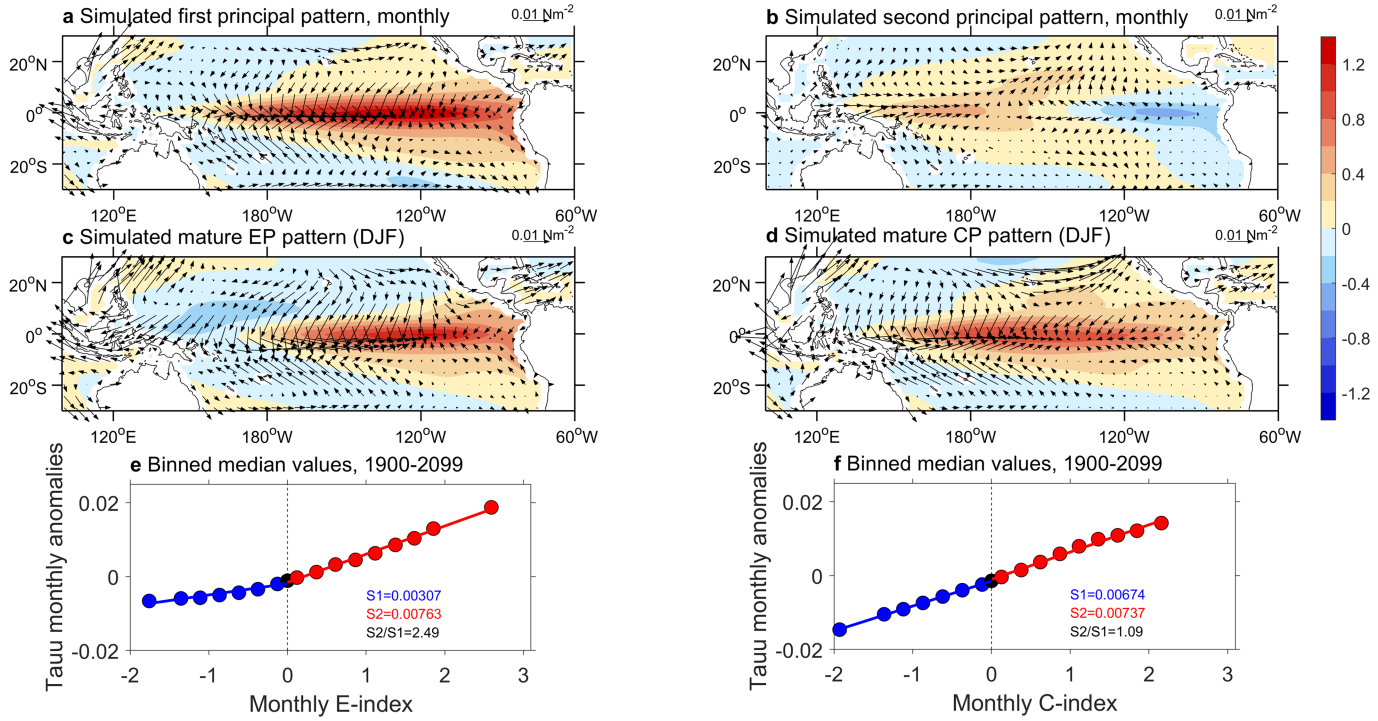
Extended Data Fig. 1 | Properties of the observed ENSO diversity, the associated CP and EP regimes, and the nonlinear Bjerknes feedback. **a**, **b**, The diversity means that the pattern of any ENSO event may be reconstructed by a combination of the first (**a**) and second (**b**) principal pattern from an EOF analysis on monthly SST anomalies (colour scale) and the associated wind-stress vectors (scale shown top right). The associated monthly PC time series are used to describe their evolution, and the CP- and EP-ENSO regimes by the C-index ($(PC1 + PC2)/\sqrt{2}$) and E-index ($(PC1-PC2)/\sqrt{2}$), respectively. **c**, **d**, The anomaly pattern associated with the EP-ENSO (**c**) and CP-ENSO (**d**) for December–February (DJF), the season in which ENSO events typically mature. **e**, **f**, Response to the E-index (**e**) or C-index (**f**) of monthly zonal wind-stress (Tauu) anomalies (in units of N m^{-2}) at the anomaly centre (see Methods) associated with the E- or C-index, respectively. The monthly wind-stress anomalies were binned in 0.25-s.d. E- or C-index intervals, and the median wind-stress anomaly and index are identified for each bin (circles). A separate linear regression was carried out for positive (red) and negative (blue) median index values. The ratio of the slope for the positive indices (S2) over that for the negative indices (S1) is taken as an indication of the nonlinear Bjerknes feedback, which operates in the EP-ENSO.

© 2018 Springer Nature Limited. All rights reserved.



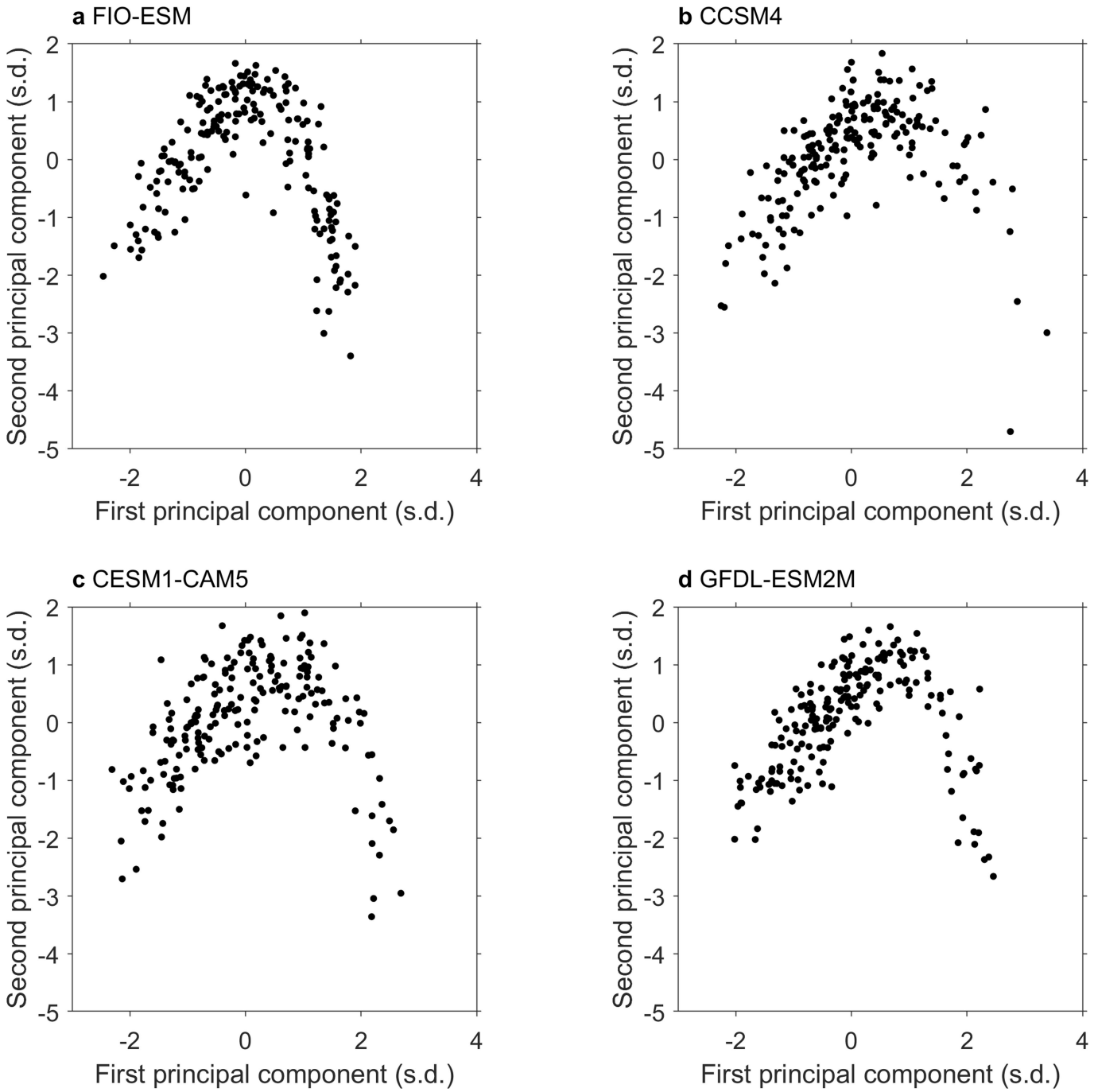
Extended Data Fig. 2 | Inter-model relationship between α and the zonal wind response to SST. a, Relationship between α and the response of monthly zonal wind anomalies to positive E-index values. Zonal wind anomalies are taken at the anomaly centre associated with the E-index. **b**, Relationship between α and the response of zonal wind anomalies to negative C-index values. Zonal wind anomalies are taken at the anomaly centre associated with the C-index.

© 2018 Springer Nature Limited. All rights reserved.



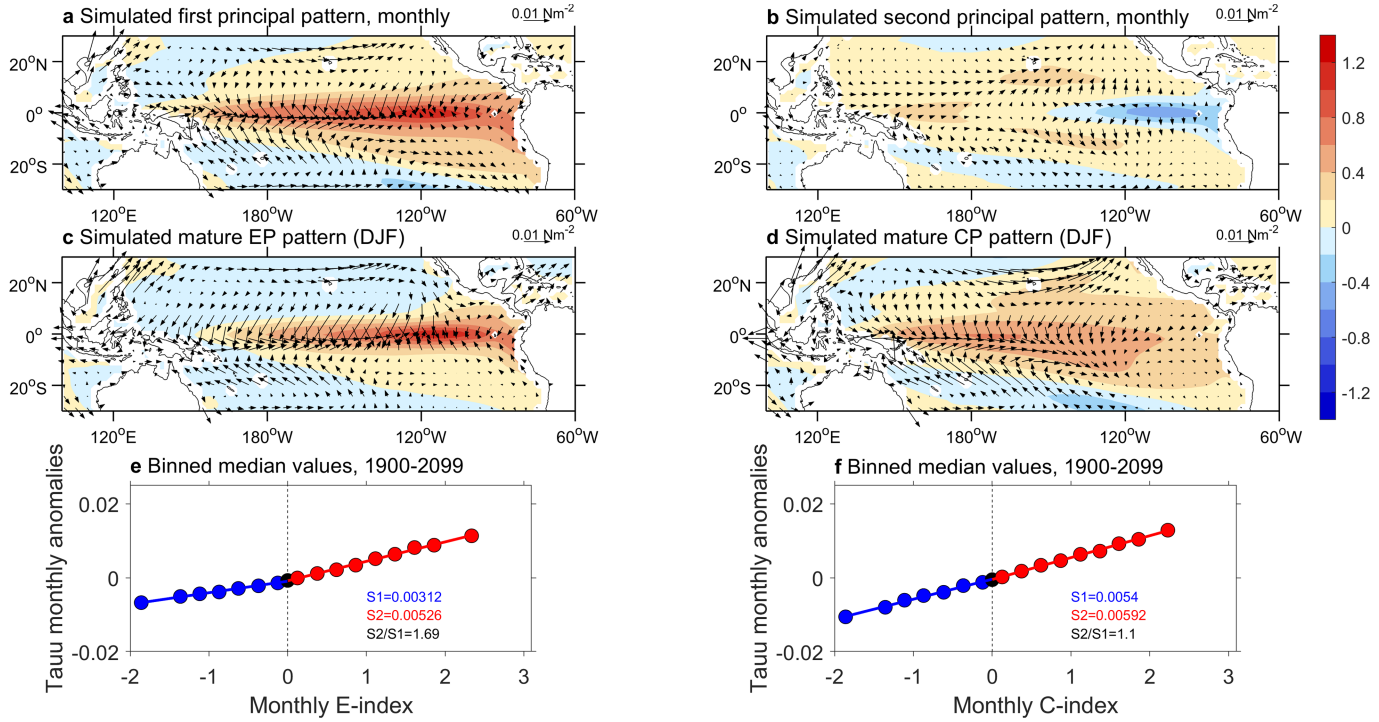
Extended Data Fig. 3 | Properties of the selected models in terms of ENSO diversity, the associated CP and EP regimes, and the nonlinear Bjerknes feedback. As in Extended Data Fig. 1, but for only the 17 selected models.

© 2018 Springer Nature Limited. All rights reserved.



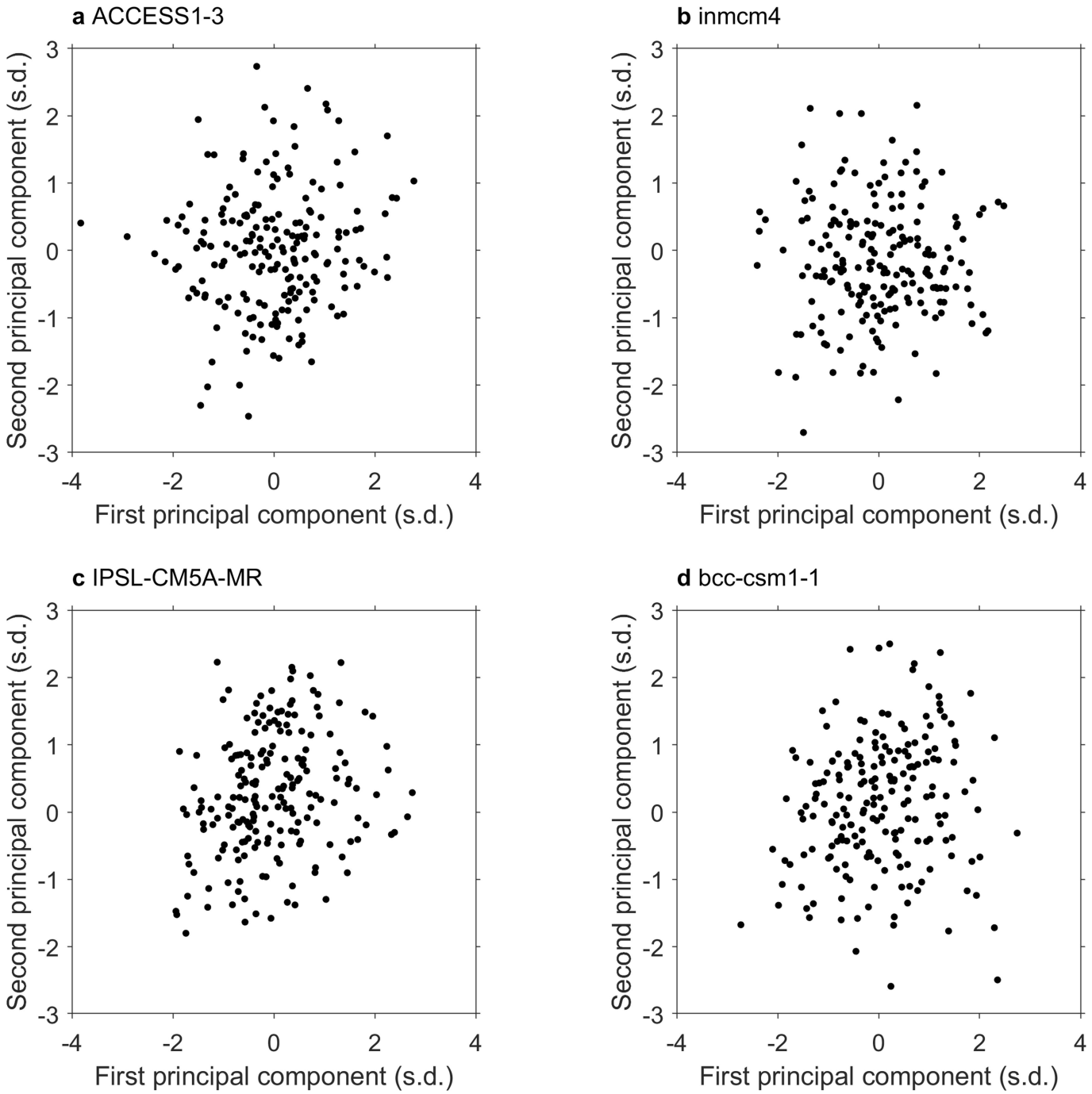
Extended Data Fig. 4 | Examples of the nonlinear relationship between the PC1 and PC2 time series in some selected models. a–d, December–February averages, with an apparent inverted V-shaped nonlinear relationship between PC1 and PC2 for FIO-ESM (**a**), CCSM4 (**b**), CESM1-CAM5 (**c**) and GFDL-ESM2M (**d**).

© 2018 Springer Nature Limited. All rights reserved.



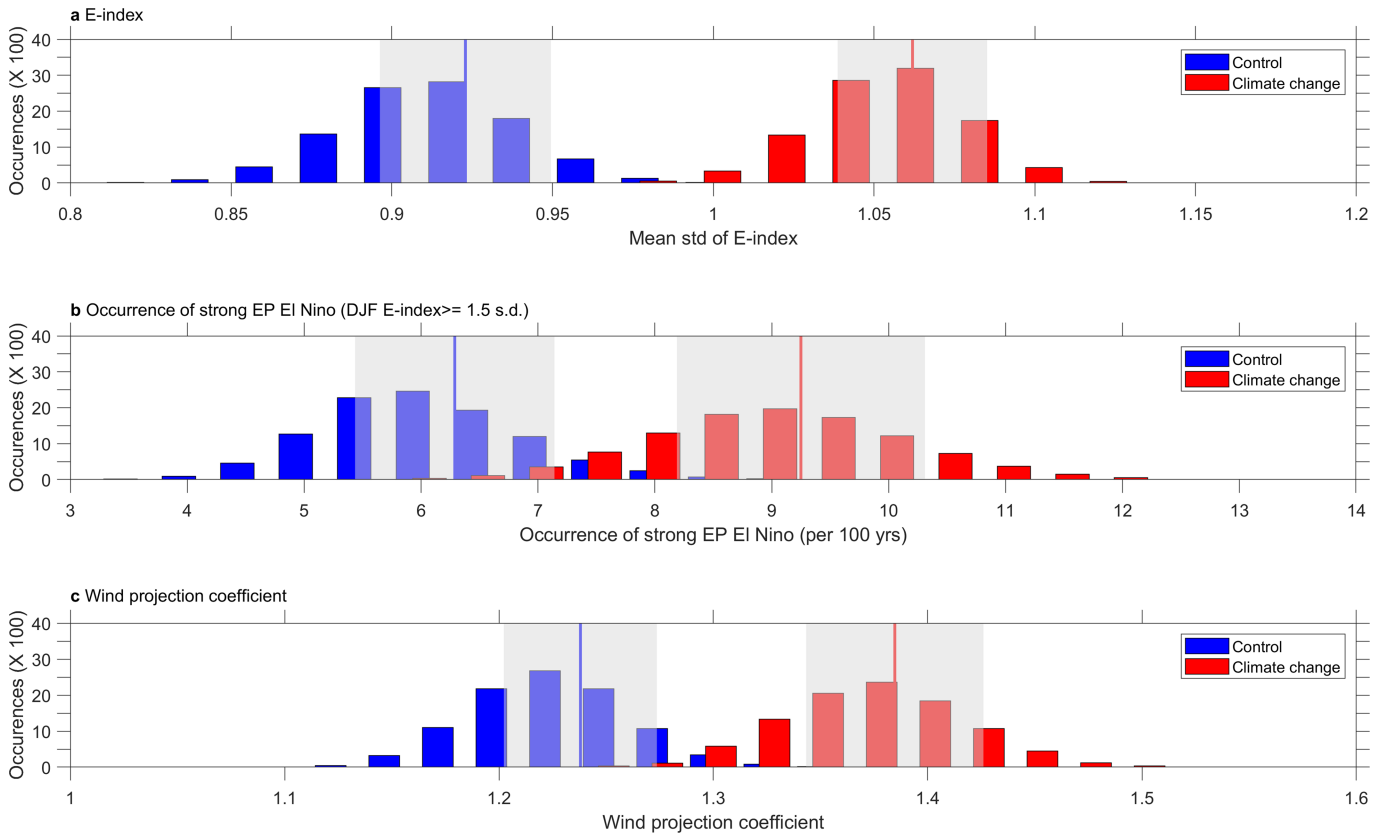
Extended Data Fig. 5 | Properties of the non-selected models in terms of ENSO diversity, the associated CP and EP regimes, and nonlinear Bjerknes feedback. As in Extended Data Fig. 3, but for only the 17 non-selected models. In this case, the nonlinear Bjerknes feedback is much weaker.

© 2018 Springer Nature Limited. All rights reserved.



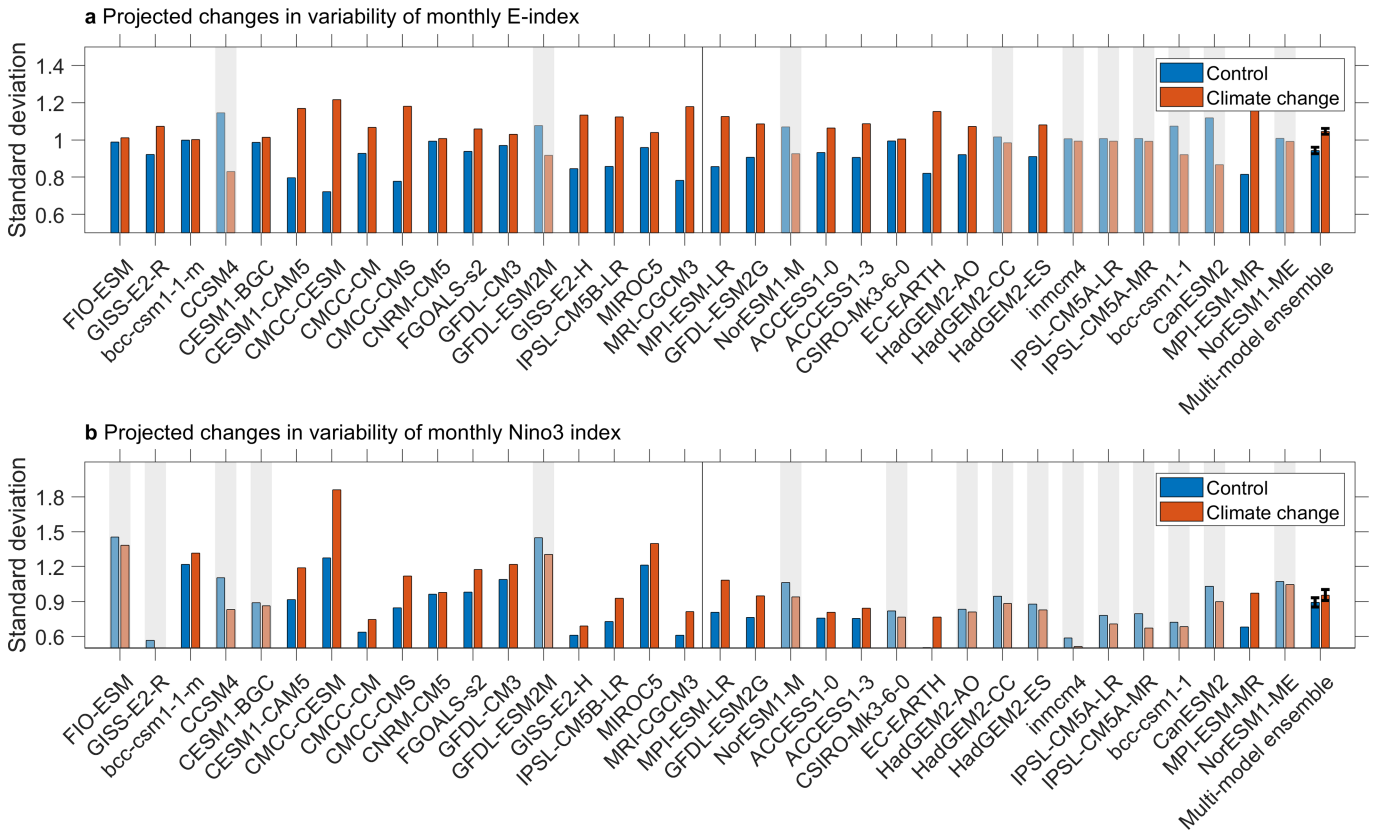
Extended Data Fig. 6 | Examples of the nonlinear relationship between the PC1 and PC2 time series in some non-selected models. **a**–**d**, December–February averages for ACCESS1-3 (**a**), inmcm4 (**b**), IPSL-CM5A-MR (**c**) and bcc-csm1-1 (**d**). In contrast to the selected models (Extended Data Fig. 4), these models display a weak or no nonlinear relationship between PC1 and PC2.

© 2018 Springer Nature Limited. All rights reserved.



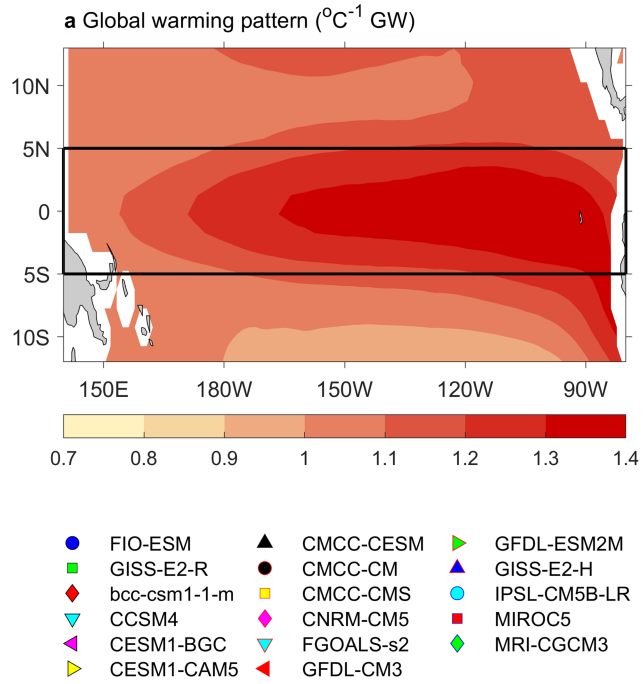
Extended Data Fig. 7 | Histograms of 10,000 realizations of a bootstrap method for the present-day (control) and future (climate change) periods. Each realization is averaged over 17 models, independently resampled randomly from the 17 selected models. The standard deviation of the 10,000 inter-realization is calculated for each period. **a**, For the E-index, the standard deviations are 0.0263 (blue) and 0.0234 (red) for the two periods. **b**, For occurrences with E-index > 1.5 s.d., the standard deviations are 0.87 (blue) and 1.06 (red) for the two periods. **c**, For the wind-projection coefficient, the standard deviations are 0.036 (blue) and 0.042 (red) for the two periods. The difference between the future and the present-day periods is greater than the sum of the two inter-realization standard deviation values (each indicated by half of the grey shaded region). The blue and red vertical lines indicate the mean values of 10,000 inter-realizations for the present-day and future periods, respectively.

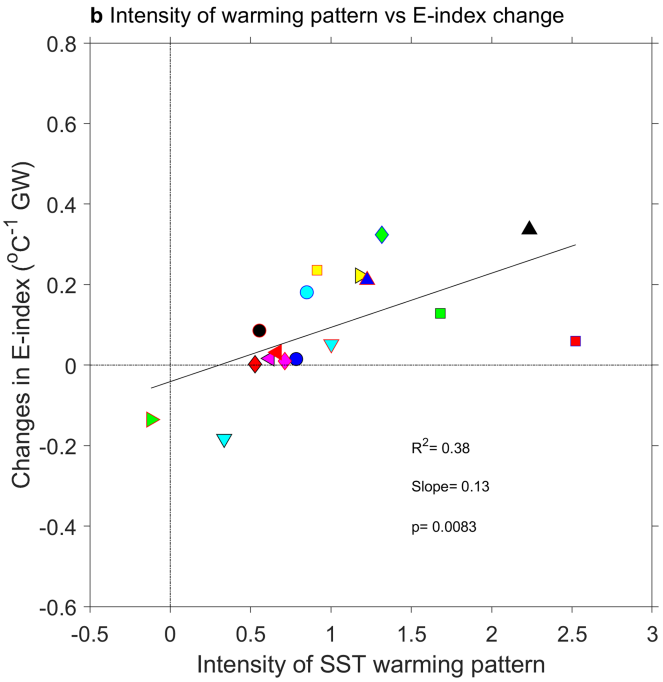
© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Projected change in EP-ENSO variability using the E-index and the Niño3 SST index. a, Comparison of the standard deviation of the E-index in the present-day (1900–1999) and future (2000–2099) 100-year periods for all 34 models. 24 of the 34 models show an increase in variance (the other 10 are greyed out). **b**, The same as **a**, but for the Niño3 SST index. Error bars in the multi-model mean are calculated as the standard deviation of the 10,000 inter-realizations. The multi-model-mean change in the E-index variance (**a**) is statistically significant at more than the 95% confidence level, but that in the Niño3 SST index is not significant (**b**). The vertical line separates the selected (left) from the non-selected (right) models.

© 2018 Springer Nature Limited. All rights reserved.





Extended Data Fig. 9 | Relationship between SST warming and change in E-index for selected models. a, Multi-model-mean warming pattern (in °C per °C of global warming (GW); colour scale). First, for each model we construct a warming pattern by calculating the difference between the average SST anomalies over the future (2000–2099) and present-day (1900–1999) periods. Second, we scale this difference by the increase in global-mean SST simulated by the model over the corresponding period. Finally, we take the mean of the scaled difference over all models to construct the multi-model-mean warming pattern. **b**, Inter-model relationship between the intensity of the SST warming pattern (**a**) and change in E-index, also scaled by the corresponding increase in global-mean SST in each model. The intensity of the scaled SST warming pattern for each model is obtained by regressing the scaled SST warming pattern for each model onto the scaled multi-model-mean SST warming pattern, using the region indicated by the black box in **a**. The inter-model relationship is statistically significant above the 95% confidence level, with the statistical properties shown.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Details of the 34 models

	Models	Data available	EP (Tauu)	CP (Tauu)	EP (SST)	CP (SST)
1	FIO-ESM	sst,τx,τy,so,T	222.25	184.75	250.75	232.75
2	GISS-E2-R	sst,τx,τy,so,T	201.25	166.75	247.75	198.25
3	bcc-csm1-1-m	sst,τx,τy,so,T	196.75	178.75	243.25	225.25
4	CCSM4	sst,τx,τy,so,T	205.75	189.25	244.75	195.25
5	CESM1-BGC	sst,τx,τy,so,T	201.25	195.25	241.75	198.25
6	CESM1-CAM5	sst,τx,τy,so,T	193.75	142.75	223.75	175.75
7	CMCC-CESM	sst,τx,τy,so,T	220.75	172.75	234.25	187.75
8	CMCC-CM	sst,τx,τy,so,T	192.25	177.25	235.75	196.75
9	CMCC-CMS	sst,τx,τy,so,T	189.25	166.75	241.75	187.75
10	CNRM-CM5	sst,τx,τy,so,T	190.75	165.25	247.75	202.75
11	FGOALS-s2	sst,τx,τy,so,T	199.75	181.75	231.25	198.25
12	GFDL-CM3	sst,τx,τy,so,T	204.25	141.25	244.75	184.75
13	GFDL-ESM2M	sst,τx,τy,so,T	196.75	142.75	246.25	189.25
14	GISS-E2-H	sst,τx,τy,so,T	223.75	141.25	247.75	207.25
15	IPSL-CM5B-LR	sst,τx,τy,so,T	216.25	165.25	252.25	202.75
16	MIROC5	sst,τx,τy,so,T	204.25	142.75	237.25	175.75
17	MRI-CGCM3	sst,τx,τy,so,T	222.25	142.75	240.25	190.75
18	MPI-ESM-LR	sst,τx,τy,so,T	142.75	142.75	243.25	175.75
19	GFDL-ESM2G	sst,τx,τy,so,T	142.75	142.75	252.25	169.75
20	NorESM1-M	sst,τx,τy,so,T	205.75	193.75	252.25	199.75
21	ACCESS1-0	sst,τx,τy	232.75	171.25	246.25	234.25
22	ACCESS1-3	sst,τx,τy,so,T	222.25	165.25	253.75	228.25
23	CSIRO-Mk3-6-0	sst,τx,τy,so,T	178.75	141.25	196.75	159.25
24	EC-EARTH	sst, so,T	NaN	NaN	252.25	201.25
25	HadGEM2-AO	sst,τx,τy	228.25	174.25	241.75	229.75
26	HadGEM2-CC	sst,τx,τy	229.75	184.75	246.25	237.25
27	HadGEM2-ES	sst,τx,τy,so,T	229.75	192.25	244.75	234.25
28	inmcm4	sst,τx,τy	217.75	192.25	240.25	192.25
29	IPSL-CM5A-LR	sst,τx,τy,so,T	205.75	198.25	258.25	201.25
30	IPSL-CM5A-MR	sst,τx,τy,so,T	198.25	187.75	258.25	193.75
31	bcc-csm1-1	sst,τx,τy,so,T	210.25	183.25	240.25	234.25
32	CanESM2	sst,τx,τy	145.75	145.75	238.75	201.25
33	MPI-ESM-MR	sst,τx,τy,so,T	163.75	141.25	238.75	177.25
34	NorESM1-ME	sst,τx,τy,so,T	207.25	195.25	249.25	204.25

The final four columns show the longitudes (° E) of the monthly maximum zonal wind-stress (Tauu) anomalies and SST anomalies in the EP and CP patterns, averaged over 5° S–5° N, for the 34 CMIP5 CGCMs, each forced under greenhouse gas concentration scenario RCP8.5. The first 17 models listed produce a reasonably large $|\alpha|$ and a nonlinear PC1–PC2 relationship (Fig. 1c, d). The zonal wind-stress anomalies associated with the EP and CP patterns are obtained by regressing monthly anomalies onto the monthly E-index and C-index, respectively; the SST anomaly pattern and centre are identified in a similar manner, except using the December–February averages of the E and C indices and SST anomalies. 'sst', 'τx', 'τy', 'so' and 'T' in the third column indicate SST, zonal wind stress, meridional wind stress, ocean salinity and temperature that are available, respectively. The 17 non-selected models are shown in grey. 'NaN' indicates data not available.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0715-9

Widespread but heterogeneous responses of Andean forests to climate change

Belén Fadrique[1], Selene Báez[2,3], Álvaro Duque[4], Agustina Malizia[5], Cecilia Blundo[5], Julieta Carilla[5], Oriana Osinaga-Acosta[5], Lucio Malizia[6], Miles Silman[7], William Farfán-Ríos[7,8], Yadvinder Malhi[9], Kenneth R. Young[10], Francisco Cuesta C.[3,11], Jurgen Homeier[12], Manuel Peralvo[3], Esteban Pinto[3], Oswaldo Jadan[13], Nikolay Aguirre[14], Zhofre Aguirre[14] & Kenneth J. Feeley[1,15]*

Global warming is forcing many species to shift their distributions upward, causing consequent changes in the compositions of species that occur at specific locations. This prediction remains largely untested for tropical trees. Here we show, using a database of nearly 200 Andean forest plot inventories spread across more than 33.5° latitude (from 26.8° S to 7.1° N) and 3,000-m elevation (from 360 to 3,360 m above sea level), that tropical and subtropical tree communities are experiencing directional shifts in composition towards having greater relative abundances of species from lower, warmer elevations. Although this phenomenon of 'thermophilization' is widespread throughout the Andes, the rates of compositional change are not uniform across elevations. The observed heterogeneity in thermophilization rates is probably because of different warming rates and/or the presence of specialized tree communities at ecotones (that is, at the transitions between distinct habitats, such as at the timberline or at the base of the cloud forest). Understanding the factors that determine the directions and rates of compositional changes will enable us to better predict, and potentially mitigate, the effects of climate change on tropical forests.

As global temperatures rise, species are predicted to shift their geographical distributions towards cooler latitudes and elevations[1]. These 'species migrations' (here referring to all modes of range changes, including expansions, contractions and shifts[2,3]) have been observed in many different species and systems[2,4,5]. However, the vast majority of studies that have investigated species migrations are from temperate or boreal systems, and little information is available about the responses of tropical and subtropical species—and in particular tropical plant species—to climate change[2,6]. This is despite the fact that tropical plants may be especially susceptible to climate change because of their narrow thermal niches[7], and the fact that species migrations out of tropical lowlands can cause biotic attrition[8] and losses of local biodiversity.

Species migrations of tropical plants

For tropical and subtropical plants, only a small set of studies have researched species migrations. The most direct approach to detect species migrations is to quantify changes in the ranges of species over time, usually by tracking shifts in the mean or upper elevational range limits of species. For example, in the alpine Himalayas (>4,000 m above sea level, (m a.s.l.)), the upper range limits of nearly 90% of investigated plant species have risen since 1850[9]. In another study, 58% of plant species studied in Taiwan shifted their ranges upwards over a 100-year period[10]. In Hawaii, 67% of the studied grass species increased their maximum elevation levels over a 42-year period[11], and in Ecuador, 88% of the studied alpine plant species expanded their upper range limits to higher elevations over a 200-year period[12]. Although these species-specific studies provide compelling evidence that many tropical and subtropical plant species are shifting their ranges upslope, the approach is of limited applicability because it requires accurate maps of the ranges of individual species, or range limits, at multiple times. Long-term species-specific data are not available for the majority of tropical species. Indeed, even the current ranges of most tropical plant species remain unknown, making it impossible to test for temporal range shifts. For example, the Himalayan study included only 124 species and the Taiwanese study 24 species—small fractions of the total plant diversity in either of these areas. Thousands of other species from throughout the tropics and subtropics are similarly excluded from these types of studies because of the lack of accurate distribution data[6,13].

Another approach that enables the integration of data that are more-readily available from more areas and for more species is to analyse changes in the taxonomic or functional composition of communities over time[5]. More specifically, a community temperature index (CTI) can be used to characterize communities based on the relative abundances of species with different thermal affiliations, or optima, and to test the prediction that species migrations should cause directional changes in composition over time[14]. For example, upward species migrations will result in greater relative abundances of more-thermophilic species from relatively warmer climates at any given elevation. In other words, upward species migrations should cause increases in the CTI of communities—a phenomenon referred to as 'thermophilization'. Changes in the CTI and thermophilization have been used as evidence of latitudinal migrations of bird and butterfly species[15], as well as temperate lowland[16] and alpine plant species[17]. Changes in CTI, or other analogous indices, have also been calculated from fossil pollen records to estimate rates of plant species migrations in response to past climate change[18].

[1]Department of Biology, University of Miami, Coral Gables, FL, USA. [2]Departamento de Biología, Escuela Politécnica Nacional, Quito, Ecuador. [3]Consortium for the Sustainable Development of the Andean Ecoregion (CONDESAN), Quito, Ecuador. [4]Departamento de Ciencias Forestales, Universidad Nacional de Colombia–Sede Medellín, Medellín, Colombia. [5]Instituto de Ecología Regional (IER), Universidad Nacional de Tucumán (UNT) & Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Yerba Buena, Argentina. [6]Facultad de Ciencias Agrarias, Universidad Nacional de Jujuy, Jujuy, Argentina. [7]Department of Biology and Center for Energy, Environment and Sustainability, Wake Forest University, Winston-Salem, NC, USA. [8]Herbario Vargas (CUZ), Universidad Nacional de San Antonio Abad del Cusco, Cusco, Peru. [9]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK. [10]Department of Geography & the Environment, University of Texas at Austin, Austin, TX, USA. [11]Paleoecology & Landscape Ecology, Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, Netherlands. [12]Plant Ecology, University of Göttingen, Göttingen, Germany. [13]Universidad de Cuenca, Facultad de Ciencias Agropecuarias, Cuenca, Ecuador. [14]Programa de Investigación Biodiversidad y Servicios Ecosistémicos, Universidad Nacional de Loja, Loja, Ecuador. [15]Fairchild Tropical Botanic Garden, Coral Gables, FL, USA. *e-mail: kjfeeley@gmail.com

© 2018 Springer Nature Limited. All rights reserved.

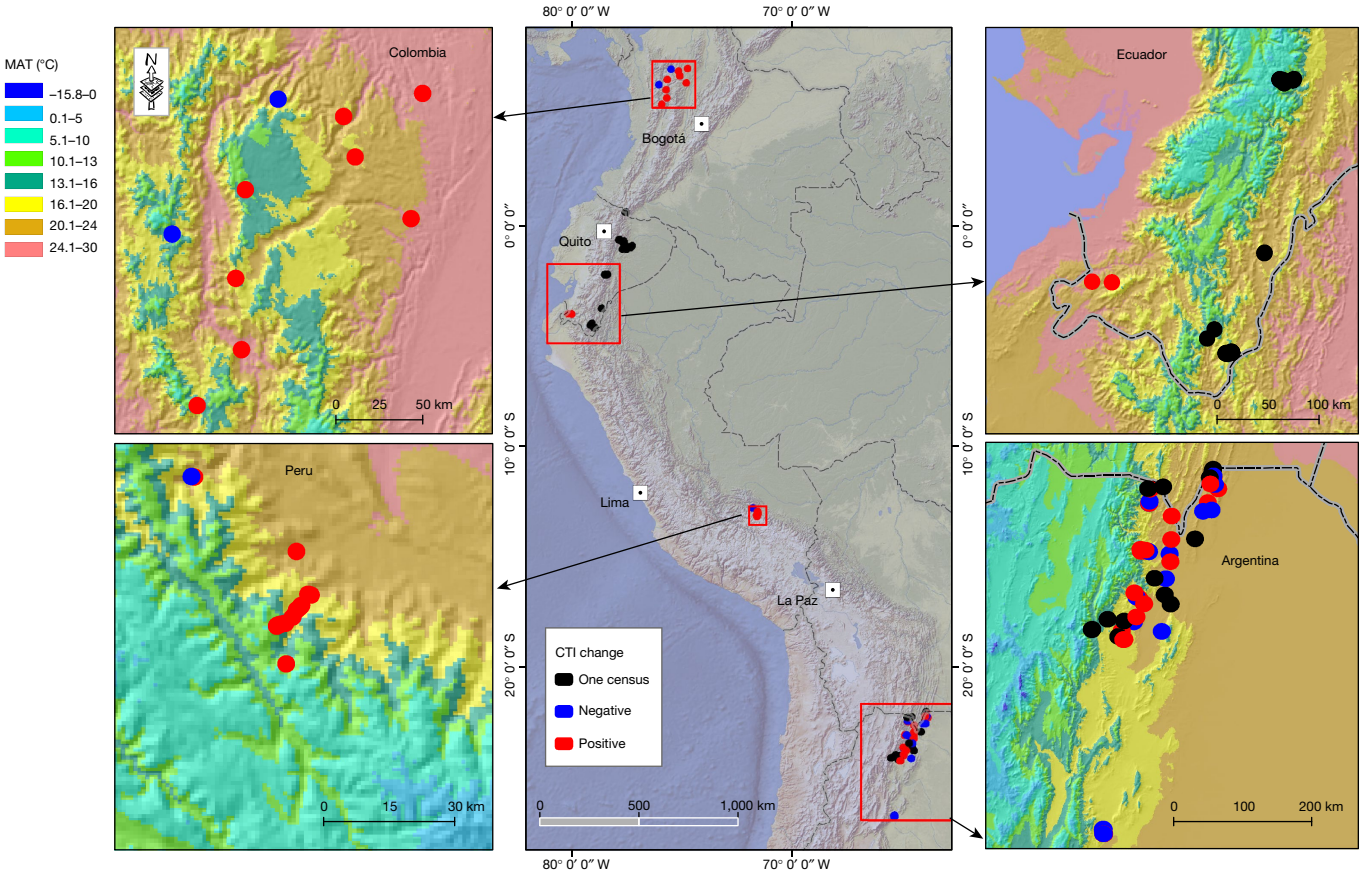


Fig. 1 | Map of Andean forest plot locations. The map shows the locations of the 186 Andean forest plots that are included in the analyses. Red points indicate plots with positive TR_{plot} (which is the annualized change in the CTI of a plot); blue points are plots with negative TR_{plot}; black points are plots with only one census and for which it was therefore not possible to calculate the TR_{plot}.

Tracking changes in the CTI is an especially useful tool for analysing the effects of climate change in hyperdiverse systems, such as tropical forests, because it does not require precise information about the range limits of individual species and enables the integration of census data from multiple locations and years. Several recent studies have used the CTI to characterize changes in the composition of tree species in tropical montane forests owing to contemporary species migrations. These studies from Peru, Costa Rica and Colombia found that focal tree communities mostly show increases in their CTI over time[19–21]. The thermophilization of these forests was hypothesized to be due primarily to increased mortality of the more heat-sensitive (that is, less-thermophilic) tree species as temperatures increased[21].

The above studies indicate that climate change is causing rapid shifts in the distributions of many tropical trees, which in turn is leading to directional changes in forest composition. However, important questions remain about the generalizability of these results. Specifically, it remains unclear how widespread and uniform the process of thermophilization is in tropical forests and what factors cause variation in thermophilization rates between different communities. In this study, we address these questions in the Tropical Andes Biodiversity Hotspot, from Colombia to Argentina in South America. We assess variability in rates of thermophilization between sites to gain insights into the factors that may be slowing, or preventing, species migrations in some areas.

Thermophilization of Andean forests

To examine temporal changes in the composition of tropical tree species at a large spatial scale, we collated a database of forest censuses from 186 inventory plots spread throughout the tropical and subtropical Andes Mountains of Colombia, Ecuador, Peru and northern Argentina[22] (Fig. 1). The plots span an elevation gradient of more than 3,000 m, corresponding to a gradient of approximately 14 °C in mean annual temperature (MAT). A total of 120 plant families, 528 genera and 2,024 tree species (including palms, tree ferns and lianas) occur in the study plots and were included in our analysis (further information about the plots is provided in Table 1 and in Supplementary Table 1). Using this dataset of tree-species composition in Andean forests, we analysed the relationships between the CTI and the environmental temperature and elevation of the plots. We then tested for thermophilization—that is, increases in CTI over time—in the plots that had been censused repeatedly ($n = 64$). We also applied a new analytical approach to the combined dataset of all plots ($n = 186$) to determine how rates of thermophilization relate to elevation and temperature across the Andes.

To look at patterns of species composition and compositional change, we first calculated the CTI (°C) for each plot during each census. CTI is the mean of the thermal optima of all species that were found in a plot weighted by their relative abundances. The thermal optimum of each species was calculated as the mean of the MATs[23] at the locations where each species are known to occur based on collection records obtained through the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/).

We next analysed the relationship between the CTI and the MAT of the plots to assess the role of regional temperatures in structuring community assembly. The average CTI of the plots ranged from 12.2 to 23.8 °C and was strongly positively correlated with MAT (slope = 0.71, $R = 0.92$, $P < 0.001$; Extended Data Fig. 1a). This indicates that the functional composition (that is, the relative abundances of species with different thermal optima) of the plots is strongly determined by temperature. In other words, plots at similar regional temperatures have similar CTI because of similar relative abundances of more- or less-thermophilic species, even if the plots are separated by as much as 4,000 km (for example, plots in Argentina versus Colombia) and have little to no taxonomic overlap in species. Given the strength of

© 2018 Springer Nature Limited. All rights reserved.

Table 1 | Description of the Andean forest plot database

Countries	Argentina	Colombia	Ecuador	Peru	Total
Number of plots	52 (48)	10 (10)	110 (6)	14 (14)	186 (78)
Total plot area in ha	55.84	10	16.72	14	96.56
Number of plots with multiple censuses	38 (34)	10 (10)	2 (2)	14 (14)	64 (60)

The number of 1-ha plots is shown in brackets.

MAT in determining the functional composition of Andean forests, global warming should manifest as temporal increases in the CTI of the study plots.

To test for thermophilization, we first looked at changes in CTI over time in all plots that had been censused more than once. We calculated the annualized rate of change in the CTI of each plot in all possible census intervals ($n = 176$ census intervals, Fig. 2) and used the rate of change in CTI between the initial and final censuses as the best estimate of the thermophilization rate of each plot (TR_{plot}). Of the 64 plots with repeated census data, 46 (72%) increased in CTI (that is, had positive thermophilization rates), indicative of increasing relative abundances of species from relatively warmer climates. The number of plots with positive TR_{plot} is more than expected under the null expectation of equal proportions of plots with positive and negative TR_{plot} values that would occur due to random fluctuations in composition over time (binomial probability <0.001). Of the 23 plots that were censused repeatedly, 43% consistently increased in CTI and had positive TR_{plot} over all intervals. By contrast, only one of the plots (4%) had negative TR_{plot} in all intervals. The mean thermophilization rate measured across all census intervals was 0.0066 °C per year (95% confidence interval = 0.004– 0.009 °C per year) (see Methods and Extended Data Fig. 2 for an alternative method of calculating TR_{plot}).

To assess how thermophilization rates relate to plot temperatures and elevations, we integrated the compositional information from all 186 inventory plots, including those that had been censused only once, and calculated a running mean of CTI per MAT in overlapping five-year census intervals between 2000 and 2015 (Fig. 3a). We then calculated the thermophilization rate (TR_{MAT}) as the slope of the linear least-square regression between its mean CTI and the midpoint of the respective time period (Fig. 3b). TR_{MAT} was significantly positive at most MATs and elevations, consistent with the widespread thermophilization observed in the per-plot analysis described above. Using a linear mixed-effect model of CTI versus year with plot identity included as a random effect ($n = 283$), we estimated that the mean TR_{MAT} was 0.003 °C per year (95% confidence interval = 0.002–0.004 °C per year). The difference between the mean TR_{MAT} and TR_{plot} is due to the inclusion of plots with single censuses and the fact that TR_{MAT} incorporates temporal changes in the CTI both within and between plots.

Ecotonal barriers to species migrations

Our results support the hypothesis that increasing temperatures are causing thermophilization of montane forests across much of the tropical and subtropical Andes. Although thermophilization is widespread, we also find that the rates of thermophilization are heterogeneous throughout the MAT and elevation gradients. Specifically, thermophilization rates were positive on average, but TR_{MAT} was negative or not significantly different from zero at the coldest and middle MATs (that is, at the highest and mid-elevations, respectively). This result mirrors patterns that have been observed within individual elevation gradients in Peru, Colombia and Costa Rica[19–21]; in all three of these gradients, the lowest thermophilization rates occurred at mid- and high elevations.

Although the tropical Andes have been identified as a 'warming hotspot'[24], some studies indicate that the warming rates vary between elevations[25]. As such, one possible explanation for the absence of significant thermophilization at high and mid-elevations is that warming rates may be slower at these elevations. Indeed, when we compare TR_{plot} to the estimated mean warming rates at each plot location (overall mean warming = +0.06 °C per year since 1990), we find that there is an overall positive correlation ($R = 0.30$, $P < 0.01$) and that TR_{plot} is negative at six out of the seven sites at which temperatures decreased (Fig. 4). We also find a generally positive relationship between warming rates and TR_{MAT} (Extended Data Fig. 3). Although these analyses suggest that differences in regional warming rates may be contributing to variation in thermophilization rates, the relationships are fairly weak and it is probable that other factors—as discussed below—are also important in determining rates of compositional change in Andean forests.

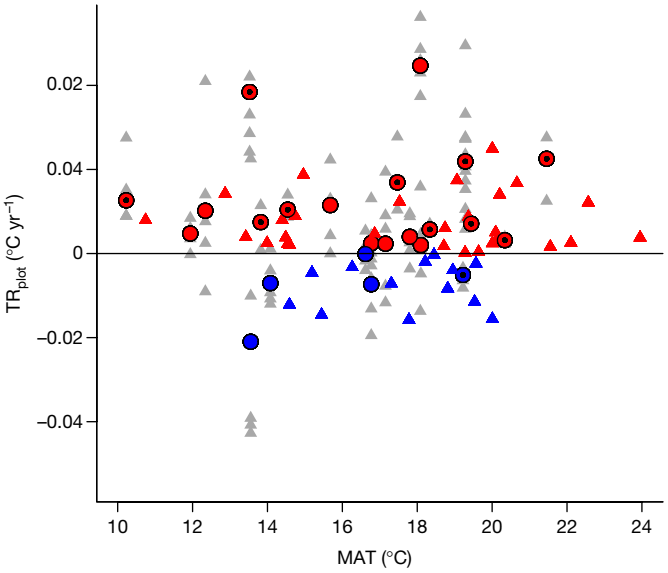


Fig. 2 | Thermophilization rates of repeatedly censused plots. TR_{plot} and MAT values were calculated for the Andean forest plots with multiple censuses. $n = 64$. Triangular grey points represent the annual change in the CTI for each of the possible census intervals ($TR_{interval}$); coloured triangular points represent plots with only two censuses and therefore only one interval. Circular points represent the average annual change in the CTI over the complete study period for plots with more than two censuses (that is, the annualized difference between final and initial CTI). Positive and negative thermophilization rates are coloured red and blue, respectively. Circles with black centres indicate plots for which the CTI changed consistently in one direction (that is, positive or negative) across all intervals.

An alternative or additional factor that may be driving differences in thermophilization rates across elevations is the presence of several distinct ecotones along the slopes of the Andes; for example, the transition from montane rainforest to cloud forest (that is, the cloud base) at mid-elevations and the transition from closed-canopy forest to open alpine grasslands (that is, the timberline) at high elevations. Conditions at ecotones can be biotically and abiotically distinct from surrounding forests, potentially reducing establishment success of colonizers and favouring stability of incumbent communities. For example, the cloud-base ecotone represents an inflexion point in many environmental variables such as precipitation, diurnal temperature range[26], soil water content[27] and light availability[28]. If the tree communities in and around ecotones are comprised predominantly of specialist species able to cope with the unique conditions that occur at these sites, then it may be harder for the composition to change, because any change will require the encroachment of non-specialized species. As one conceptual example, many high-elevation forests are dominated by just a single to few species of trees that are specifically adapted to the unique environmental conditions that occur near the timberline. Even if rising temperatures cause decreased reproduction, performance and/or increased mortality of these species (or even the complete loss of some species), changes in the CTI will be small because all of the remaining species will have similar thermal optima. In cases such as these, thermophilization can only occur if a species from lower elevations (that is, with higher thermal optima) expands its range and recruits into the ecotonal forest. In

© 2018 Springer Nature Limited. All rights reserved.

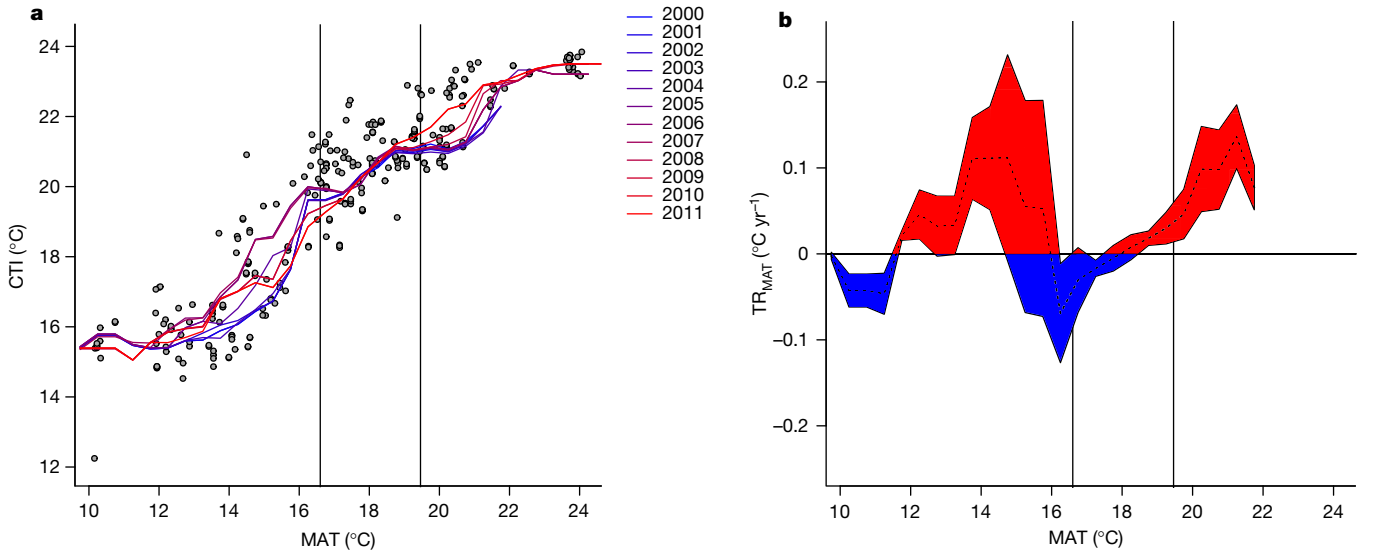


Fig. 3 | Thermophilization rates of Andean forest plots. a, The CTI and MAT values for all plot censuses are shown as points. $n = 283$. The lines indicate the mean CTI compared to MAT, calculated for overlapping five-year time periods from 2000 to 2015. **b**, The thermophilization rates for thermal bands (TR_{MAT}; the annualized change in the mean CTI of all plots within a thermal band) was compared to MAT values. $n = 283$ plot censuses, assigned to 28 thermal bands. The dashed line indicates the mean TR_{MAT} and the coloured area indicates the 95% confidence interval of TR_{MAT} with positive and negative values coloured red and blue, respectively. The black rectangle encompasses the approximate MATs encompassing cloud base in each country (Argentina, 1,000 m a.s.l.; Ecuador, 1,400 m a.s.l.; Colombia, 1,500 m a.s.l.; Peru, 1,600 m a.s.l.). Timberline occurs at approximately 5–7 °C MAT.

support of this hypothesis, we find that (1) plots with negative thermophilization rates have lower species richness than expected based on their MAT (Extended Data Fig. 4), (2) more-specialized communities have slower rates of thermophilization as indicated by a positive correlation between measures of intraspecific variation in the thermal optima of co-occurring species and the corresponding TR_{plot} (Extended Data Fig. 5) and (3) the absolute abundance (basal area) of more-thermophilic species (that is, with thermal optima higher than the CTI of the plot) remained stable or decreased in low- and mid-temperature plots, but generally increased in high-temperature plots (by contrast, the absolute abundance of less-thermophilic species with thermal optima below the CTI of the plot increased at mid-temperatures) (Extended Data Fig. 6). In other words, slow or negative thermophilization rates are associated with areas in which warming rates are slower, and/or with lower-diversity and more-specialized forests near ecotones in which there is little ingrowth or recruitment of more-thermophilic species.

In addition, variation in thermophilization rates can increase if the ranges of some species are limited by biotic interactions[29], non-climatic factors (for example, topography or soil-nutrient composition) or climatic factors that do not change concomitantly with temperature (for example, cloud cover or water availability)[30]. For example, if water availability decreases with elevation, then changes in precipitation and rising temperatures could cause drought-sensitive species to migrate downslope[31], resulting in negative TR_{plot} estimates. Similarly, changes in other environmental constraints such as light exposure or the frequency of frost events can change in unexpected and nonlinear ways (for example, owing to changes in cloud cover or 'cloud lifting'), potentially leading to downward migrations of some species in some areas. For example, there is evidence that the upper range limits of some Andean tree species are set by cold night-time temperatures and frost events. Despite rising MATs, the frequency and magnitude of frost events is increasing in some areas; this could prevent upward migrations and potentially cause the downward migrations of some species[32].

Finally, we cannot rule out the influence of idiosyncratic processes or events on community composition and thus thermophilization rates[33]—especially given the relatively short duration of this study in relation to the lifespans of most trees. In particular, the two plots with the most-negative TR_{plot} values may have been influenced by site-specific factors including increased understory recruitment due to reduced herbivory[34] and high growth rates of less-thermophilic understory species during certain years[35]. For this study, we focused exclusively on the effects of rising temperatures on tree species composition, but multiple forces (both climatic and non-climatic) can undeniably affect the suitability of habitats for different species[36] and, therefore, the CTI[33]; uncovering and integrating these other factors must be a priority of future studies.

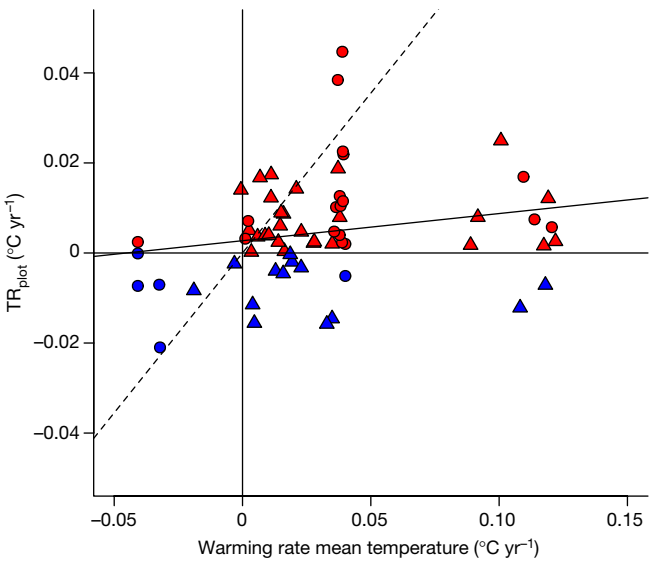


Fig. 4 | Mean warming rate at plots. The TR_{plot} was compared to the estimated warming rate (average annual change in mean temperature between 1990 and 2013) at each of the plot locations. $n = 64$ plots, Spearman correlation two-sided $R = 0.30$, 95% confidence interval = 0.06–0.50, $P < 0.01$; solid diagonal line. The triangular points are plots with only two censuses and the circular points represent plots with more than two censuses. Plots with positive and negative TR_{plot} are coloured red and blue, respectively. The dashed line shows the expected relationship between TR_{plot} and warming rate based on the observed relationship between CTI and MAT ($TR_{plot} = 0.71 \times \Delta MAT$).

© 2018 Springer Nature Limited. All rights reserved.

Discussion

Our analyses indicate widespread thermophilization but with rates of compositional change that vary across elevation—potentially owing to differences in warming rates, the occurrence of ecotonal 'roadblocks' and/or the influence of factors other than temperature in setting the range limits of some species. Although we are confident that these findings are robust, we acknowledge two limitations of this study. First, the data used in our analyses come from a single, albeit extremely important, region of the tropics—the Tropical Andes Biodiversity Hotspot[37]—and it remains uncertain how other tropical forests and ecosystems are responding to climate change. Although comparable studies are clearly needed for other tropical and subtropical regions, there is good reason to suspect that other forests are undergoing similar changes in composition. As discussed above, studies from other parts of the tropics have all shown evidence of species migrations and thermophilization of plant communities[9–12,19–21]. Similarly, studies from various regions of the tropics have shown evidence of upward migration of animal species (for example, birds, insects and herpetofauna) and communities[38–40]. Second, although the observed shifts in the composition of tropical and subtropical Andean forests towards having greater relative abundances of thermophilic species is consistent with upward species migrations, these data alone cannot be used to determine which species are migrating or the specific manner in which the ranges of individual species are changing over time (for example, by range expansion, contraction or shifts)[2]. To help to resolve these questions, species-specific analyses looking at population demographics and range dynamics are required. In addition, experimental studies will be crucial for determining the specific way(s) in which changes in different climatic factors are affecting individual species and the consequences for ecosystem processes and services[41].

Despite these limitations, this study provides comprehensive evidence that many tropical and subtropical forests are changing directionally in composition over time, most probably as a response to global warming. It is troubling to note that in all but a few plots, rates of compositional change are markedly slower than regional warming (Fig. 4). Indeed, given that global temperatures have been rising for over a century, the 'slow' rates of compositional change that are observed here (on average 10 times slower than changes in regional MAT) suggest that many tropical tree species may already be occurring in sub-optimal conditions. The disequilibrium between rates of compositional and climate change[42], together with potential ecotonal barriers to species migrations[30], raises concerns regarding the future of tropical montane forests and the many important ecosystem services that they provide. Andean forests must be added to the growing list of ecosystems and species that lack the ability to quickly and cohesively respond to climate change[15,16,43] and thus face high risk of extinction, biodiversity loss and functional collapse[44]. Modelling and conservation efforts must account for compositional lags and prepare for the likelihood that as global warming continues to accelerate, tropical forests will fall even further behind.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0715-9

Received: 17 November 2017; Accepted: 17 September 2018;
Published online 14 November 2018.

1. Thuiller, W. Climate change and the ecologist. *Nature* **448**, 550–552 (2007).
2. Lenoir, J. & Svenning, J. C. Climate-related range shifts - a global multidimensional synthesis and new research directions. *Ecography* **38**, 15–28 (2015).
3. Feeley, K. J., Rehm, E. M. & Machovina, B. The responses of tropical forest species to global climate change: acclimate, adapt, migrate, or go extinct? *Front. Biogeogr.* **4**, 69–84 (2012).
4. Chen, I. C., Hill, J. K., Ohlemüller, R., Roy, D. B. & Thomas, C. D. Rapid range shifts of species associated with high levels of climate warming. *Science* **333**, 1024–1026 (2011).
5. Steinbauer, M. J. et al. Accelerated increase in plant species richness on mountain summits is linked to warming. *Nature* **556**, 231–234 (2018).
6. Feeley, K. J., Silman, M. R. & Duque, A. Where are the tropical plants? A call for better inclusion of tropical plants in studies investigating and predicting the effects of climate change. *Front. Biogeogr.* **7**, 174–176 (2015).
7. Perez, T. M., Stroud, J. T. & Feeley, K. J. Thermal trouble in the tropics. *Science* **351**, 1392–1393 (2016).
8. Colwell, R. K., Brehm, G., Cardelús, C. L., Gilman, A. C. & Longino, J. T. Global warming, elevational range shifts, and lowland biotic attrition in the wet tropics. *Science* **322**, 258–261 (2008).
9. Telwala, Y., Brook, B. W., Manish, K. & Pandit, M. K. Climate-induced elevational range shifts and increase in plant species richness in a Himalayan biodiversity epicentre. *PLoS ONE* **8**, e57103 (2013).
10. Jump, A. S., Huang, T.-J. & Chou, C.-H. Rapid altitudinal migration of mountain plants in Taiwan and its implications for high altitude biodiversity. *Ecography* **35**, 204–210 (2012).
11. Angelo, C. L. & Daehler, C. C. Upward expansion of fire-adapted grasses along a warming tropical elevation gradient. *Ecography* **36**, 551–559 (2013).
12. Morueta-Holme, N. et al. Strong upslope shifts in Chimborazo's vegetation over two centuries since Humboldt. *Proc. Natl Acad. Sci. USA* **112**, 12741–12745 (2015).
13. Feeley, K. J. & Silman, M. R. Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Divers. Distrib.* **17**, 1132–1140 (2011).
14. Stuart-Smith, R. D., Edgar, G. J., Barrett, N. S., Kininmonth, S. J. & Bates, A. E. Thermal biases and vulnerability to warming in the world's marine fauna. *Nature* **528**, 88–92 (2015).
15. Devictor, V. et al. Differences in the climatic debts of birds and butterflies at a continental scale. *Nat. Clim. Change* **2**, 121–124 (2012).
16. Bertrand, R. et al. Changes in plant community composition lag behind climate warming in lowland forests. *Nature* **479**, 517–520 (2011).
17. Gottfried, M. et al. Continent-wide response of mountain vegetation to climate change. *Nat. Clim. Change* **2**, 111–115 (2012).
18. Bush, M. B., Silman, M. R. & Urrego, D. H. 48,000 years of climate and forest change in a biodiversity hot spot. *Science* **303**, 827–829 (2004).
19. Feeley, K. J. et al. Upslope migration of Andean trees. *J. Biogeogr.* **38**, 783–791 (2011).
20. Feeley, K. J., Hurtado, J., Saatchi, S., Silman, M. R. & Clark, D. B. Compositional shifts in Costa Rican forests due to climate-driven species migrations. *Glob. Change Biol.* **19**, 3472–3480 (2013).
21. Duque, A., Stevenson, P. R. & Feeley, K. J. Thermophilization of adult and juvenile tree communities in the northern tropical Andes. *Proc. Natl Acad. Sci. USA* **112**, 10744–10749 (2015).
22. Báez, S. et al. Large-scale patterns of turnover and basal area change in Andean forests. *PLoS ONE* **10**, e0126594 (2015).
23. Karger, D. N. et al. Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).
24. Russell, A. M., Gnanadesikan, A. & Zaitchik, B. Are the Central Andes mountains a warming hot spot? *J. Clim.* **30**, 3589–3608 (2017).
25. Vuille, M., Franquist, E., Garreaud, R., Lavado Casimiro, W. S. & Cáceres, B. Impact of the global warming hiatus on Andean temperature. *J. Geophys. Res. Atmos.* **120**, 3745–3757 (2015).
26. Rapp, J. M. & Silman, M. R. Diurnal, seasonal, and altitudinal trends in microclimate across a tropical montane cloud forest. *Clim. Res.* **55**, 17–32 (2012).
27. Palin, O. F. et al. Termite diversity along an Amazon–Andes elevation gradient, Peru. *Biotropica* **43**, 100–107 (2011).
28. Fyllas, N. M. et al. Solar radiation and functional traits explain the decline of forest primary productivity along a tropical elevation gradient. *Ecol. Lett.* **20**, 730–740 (2017).
29. Alexander, J. M., Diez, J. M. & Levine, J. M. Novel competitors shape species' responses to climate change. *Nature* **525**, 515–518 (2015).
30. Rehm, E. & Feeley, K. J. Many species risk mountain top extinction long before they reach the top. *Front. Biogeogr.* **8**, e27788 (2016).
31. Crimmins, S. M., Dobrowski, S. Z., Greenberg, J. A., Abatzoglou, J. T. & Mynsberge, A. R. Changes in climatic water balance drive downhill shifts in plant species' optimum elevations. *Science* **331**, 324–327 (2011).
32. Rehm, E. M. & Feeley, K. J. Freezing temperatures as a limit to forest recruitment above tropical Andean treelines. *Ecology* **96**, 1856–1865 (2015).
33. Bowler, D. & Böhning-Gaese, K. Improving the community-temperature index as a climate change indicator. *PLoS ONE* **12**, e0184275 (2017).
34. Malizia, A., Easdale, T. A. & Grau, H. R. Rapid structural and compositional change in an old-growth subtropical forest: using plant traits to identify probable drivers. *PLoS ONE* **8**, e73546 (2013).
35. Carilla, J. & Grau, R. Tendencias sucesionales de los bosques montanos subtropicales del noroeste argentino. *Bosque (Valdivia)* **32**, 97–111 (2011).
36. McCain, C. M. & Colwell, R. K. Assessing the threat to montane biodiversity from discordant shifts in temperature and precipitation in a changing climate. *Ecol. Lett.* **14**, 1236–1245 (2011).
37. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
38. Raxworthy, C. J. et al. Extinction vulnerability of tropical montane endemism from warming and upslope displacement: a preliminary appraisal for the highest massif in Madagascar. *Glob. Change Biol.* **14**, 1703–1720 (2008).

© 2018 Springer Nature Limited. All rights reserved.

39. Chen, I.-C. et al. Elevation increases in moth assemblages over 42 years on a tropical mountain. *Proc. Natl Acad. Sci. USA* **106**, 1479–1483 (2009).
40. Forero-Medina, G., Terborgh, J., Socolar, S. J. & Pimm, S. L. Elevational ranges of birds on a tropical montane gradient lag behind warming temperatures. *PLoS ONE* **6**, e28535 (2011).
41. Feeley, K. J. Moving forward with species distributions. *Am. J. Bot.* **102**, 173–175 (2015).
42. Svenning, J. C. & Sandel, B. Disequilibrium vegetation dynamics under future climate change. *Am. J. Bot.* **100**, 1266–1286 (2013).
43. Menéndez, R. et al. Species richness changes lag behind climate change. *Proc. R. Soc. B* **273**, 1465–1470 (2006).
44. Bertrand, R. et al. Ecological constraints increase the climatic debt in forests. *Nat. Commun.* **7**, 12643 (2016).

Acknowledgements We thank the many individuals and institutions (including the Red de Bosques Andinos, CODESAN, APECO, CONICET and RAINFOR) who are working to protect and understand Andean forests; GBIF and contributing institutions for making collection data publicly available and E. Ortíz for creating the map of plot locations. B.F. and K.J.F. were supported by the US NSF (DEB-1350125) and the Swiss Agency for Development and Cooperation. M.R.S. and W.F.R. were supported by the US NSF (DEB-1754647, DEB-1258112, and EAR-1338694). J.H. was supported by DFG Grants HO3296/2 and HO3296/4. Peruvian plot monitoring was supported by the Blue Moon Fund and the Gordon and Betty Moore Foundation's Andes to Amazon Program and RAINFOR grant 1656 (coordinated by O. Phillips). A complete list of acknowledgments and funding sources can be found in the Supplementary Information.

Reviewer information *Nature* thanks A. M. Latimer, J. Lenoir, H. Pauli and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.F. and K.J.F. designed the study, carried out the analysis and wrote the Article. S.B., A.D., A.M., C.B., J.C., O.O.-A., L.M., M.S., W.F.-R., Y.M., K.R.Y., F.C.C., J.H., M.P., E.P., O.J., N.A. and Z.A. provided data and assistance with writing.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0715-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0715-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.J.F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Data. We collated census data from 186 Andean forest inventory plots (http://redbosques.condesan.org/). Plots were originally established as parts of seven independent projects with differing motivations and methods but with common core data on the identity and size of all trees (including palms, tree ferns and lianas) that were found within each plot (first census in 1991). A subset of 64 plots had been censused repeatedly (median number of censuses = 2, maximum number of censuses = 6) providing additional data on temporal changes in species composition. In collaboration with plot managers, the collated database was cleaned and corroborated to maximize accuracy of plot metadata, species identifications and stem diameter measurements. Plots with observations that indicated secondary forest composition were excluded from the database; however, two of the included plots from Argentina showed some signs of successional processes[35] or recovery from past disturbances (cattle)[34] that may affect their understory composition (excluding these plots had no observable effects on the results). Plot elevations were estimated based on their coordinates and the SRTM 1 ArcSec Global V3 (https://lta.cr.usgs.gov) 30-m-resolution digital elevation model (DEM). Plot elevations ranged from 360 to 3,360 m a.s.l., corresponding to a MAT gradient from 24.1 to 10.2 °C. The MAT of each plot was estimated by extracting the CHELSA BIO1 values[23] (30-arcsec resolution; approximately 1 km at the equator) at the plot locations. We subsequently down-scaled these estimates to a resolution of 30 m by applying a geographically weighted regression (GWR) model[45]. For the GWR model, we used the environmental and climate data from a total of 745,878 pixels. These included the pixels that contained each of the study plots, all pixels within a 100-m radius around each of the plots, and 20,000 pixels sampled randomly from across the entire Andean study area (bounding box coordinates: 83.0° W, 63.0° W, 9.9° N, 30.0° S). In the GWR, CHELSA BIO1 (that is, MAT) was disaggregated to a 30-m resolution and included as the dependent variable. Elevation, slope and aspect[46], derived from the 30-m DEM (topographic variables calculated using the Raster package in R[47]), were set as the independent variables. Bandwidth of the GWR was set automatically based on preliminary analyses with 100,000 sample points. The relationship between plot elevation and MAT is shown in Extended Data Fig. 1c.

The combined list of tree species from all plots was submitted to the Taxonomic Name Resolution Service (TNRS; http://tnrs.iplantcollaborative.org/) version 3.0 for homogenization and validation of species names. The processing mode was 'name resolution' and the selected sources were The Plant List[48], the Global Compositae Checklist[49], the International Legume Database and Information Service[50], Tropicos[51] and USDA's Plants Database[52]. The family classification was based on Tropicos. The match accuracy threshold was set to 0.05 with partial matches allowed. All species with invalid original names (for example, sp1, indet, and so on) were assigned as 'undetermined'. Any species with an unassigned TNRS-accepted name and taxonomic status of 'no opinion', 'illegitimate' or 'invalid' were manually reviewed. The proper name was added if the species name could be confirmed on The Plant List or Tropicos; if the proper species name could not be confirmed, but the genus was valid, it was assigned the genus name and a unique species identifier. All TNRS species names with taxonomic status 'accepted' but with matching scores lower than 0.9 were also manually checked and modified following the same criteria. Families and genera were changed in accordance with the new species name. If a full species name was not provided or could not be found, the genus and/or family name were kept from the original file.

CTI. Using previously established protocols[19,21], we estimated the thermal distributions of all tree species that occurred in the inventory plots based on the locations of herbarium specimens reported for these species from the tropical and subtropical Andes. More specifically, for all species found in the study plots, all available georeferenced herbarium data records from the Andean countries of Colombia, Ecuador, Peru, Bolivia and northern Argentina (latitude <30° S) were downloaded through the GBIF data portal (https://www.gbif.org/; data downloaded on 9 October 2015, https://doi.org/10.15468/dl.bmz3hf). Any records that were tagged by the GBIF as having possible coordinate issues or that had obvious georeferencing errors (for example, falling in large bodies of water or outside the Andean study region) were discarded. The MAT at the collection locations of all specimens were estimated by extracting the MAT values from the CHLESA BIO1[23]-extrapolated climate map at a spatial resolution of 30 arcsec. We did not down-scale the climate map when extracting MAT values at collection locations owing to the low resolution and potential inaccuracies of the georeferencing data. Only a single occurrence per species was retained from each climate cell. Finally, the most climatically extreme records (that is, those outside the species' central 95% quantile of MAT) for each species were discarded to help to minimize the influence of outliers or remaining georeferencing errors.

For each species represented by ≥10 observation records (n = 1,220), we estimated the thermal optimum as the mean MAT (°C) of the collection locations (we also calculated thermal optima based on median MAT values but this did not significantly change the results). This method of estimating the thermal optimum is appropriate for mountain species as their full thermal ranges can be expressed. By contrast, lowland species may have truncated thermal ranges and therefore their geographical distributions may not provide accurate estimates of their thermal optima[53]. For species with <10 available records (n = 500) or identified at the genus level (n = 264), we estimated the thermal optimum as the average collection temperature calculated from all available records for congeneric individuals in the Andean region (changing minimum sample size criteria did not have qualitative effects on the results). Any species not identified to genus or that had insufficient records available at either the genus or species levels (n = 40) was excluded from subsequent analyses.

We calculated the CTI of each plot as the average thermal optima of the species weighted by their relative total basal area (summed cross-sectional stem area of all conspecifics measured at breast height (1.3 m above ground)) in that plot. Changes in the CTI therefore integrate the effects of tree growth, recruitment and mortality on community composition. The CTI of a plot is calculated as:

$$\mathrm{CTI} = \sum_{i=1}^{n} (\mathrm{SpOptT}_i(\mathrm{BA}_i / \mathrm{BA}_{\mathrm{plot}})) / \mathrm{BA}_{\mathrm{plot}}$$

in which n is the number of species in the focal plot, SpOptT_i is the thermal optimum for species i, and BA_i and $\mathrm{BA}_{\mathrm{plot}}$ are the basal area of species i and of the plot, respectively. All individuals available in the plot inventory datasets, regardless of minimum criteria for the diameter at breast height, were included in the analyses presented in the main text. We reran all analyses using standard criteria of including only stems with a diameter at breast height of ≥10 cm; results of these analyses are shown in Extended Data Table 1. The relationships between the plot CTI and MAT and between the plot CTI and elevation are shown in Extended Data Fig. 1a, b.

$\mathrm{TR}_{\mathrm{plot}}$. To test for changes in the species composition of the study plots over time, we calculated the annualized difference in CTI between all possible censuses for each plot that was censused more than once ($\mathrm{TR}_{\mathrm{interval}}$). The rate of change in CTI between the initial and final census was used as the best estimate of the $\mathrm{TR}_{\mathrm{plot}}$ of each plot. We then calculated the mean $\mathrm{TR}_{\mathrm{plot}}$ using the generalized linear model of $\mathrm{TR}_{\mathrm{interval}}$ (CTI change per interval, n = 176) with plot identity included as a random effect. We used a binomial probability test to determine whether the number of plots with positive $\mathrm{TR}_{\mathrm{plot}}$ values differed significantly from the null expectations of equal positive and negative changes. We also performed a Student's t-test between $\mathrm{TR}_{\mathrm{plot}}$ and a null hypothesis of no change. We repeated the above analyses using only the 61 plots with an area ≥1 ha and with ≥2 censuses and obtained nearly identical results (Extended Data Table 1). As an alternative means of calculating the overall change in the CTI of a plot over time, we also calculated $\mathrm{TR}_{\mathrm{plot}}$ as the slope of the linear least-square regression between CTI and census year. Results did not differ qualitatively from the $\mathrm{TR}_{\mathrm{plot}}$ estimates explained above (Extended Data Fig. 2).

$\mathrm{TR}_{\mathrm{MAT}}$. To integrate data from plots with single censuses and investigate how thermophilization rates vary across the MAT and elevation gradients, we analysed temporal changes in the running average of CTI versus MAT. More specifically, we divided the census data into overlapping five-year periods from 2000 to 2015 (that is, period 1 = 2000–2005, period 2 = 2001–2006, period 3 = 2002–2007…period 11 = 2010–2015). For each time period, we calculated the mean CTI of all plots that occurred within overlapping 1.5-°C thermal bands (equivalent to approximately 250 m elevation based on the regional adiabatic lapse rate) between 10 and 25 °C MAT such that band 1 = 10–11.5 °C, band 2 = 10.5–12 °C, band 3 = 11–12.5 °C… band 28 = 23.5–25.0 °C (plots were assigned to thermal bands based on their downscaled CHELSA BIO1 values (see above) such that the MAT and thermal band assignments of a plot did not change over time). To calculate the average CTI per thermal band per time period, plots were weighted by their area. For any plot censused more than once in a given time period, we used the average CTI of that plot within that period. For each thermal band with ≥10 plots, we then calculated the $\mathrm{TR}_{\mathrm{MAT}}$ as the slope of the linear least-square regression between average CTI and year (mid-point of the five-year time period). We used the same regression analyses to estimate the 95% confidence interval around the $\mathrm{TR}_{\mathrm{MAT}}$ estimates, which then allowed us to assess the significance of $\mathrm{TR}_{\mathrm{MAT}}$ at specific MATs. We calculated the mean $\mathrm{TR}_{\mathrm{MAT}}$ using the generalized linear model of CTI (n = 283) versus census year with plot identity included as a random effect.

Historic temperature change for the study area. We downloaded monthly mean temperature data at 30-arcsec resolution from 1990 to 2013 from the CHELSA Timeseries dataset (http://chelsa-climate.org/timeseries/). We extracted the information for the plot locations and calculated the annualized change in mean temperature as the slope of the linear least-square regression of temperature versus date. For plots with multiple censuses, we performed a Spearman correlation between the warming rate and $\mathrm{TR}_{\mathrm{plot}}$. We also replicated the $\mathrm{TR}_{\mathrm{MAT}}$ calculation substituting MAT with warming rate ($\mathrm{TR}_{\mathrm{warm}}$) (average thermophilization rates calculated at intervals of 0.01 °C between −0.05 and 0.15 °C).

© 2018 Springer Nature Limited. All rights reserved.

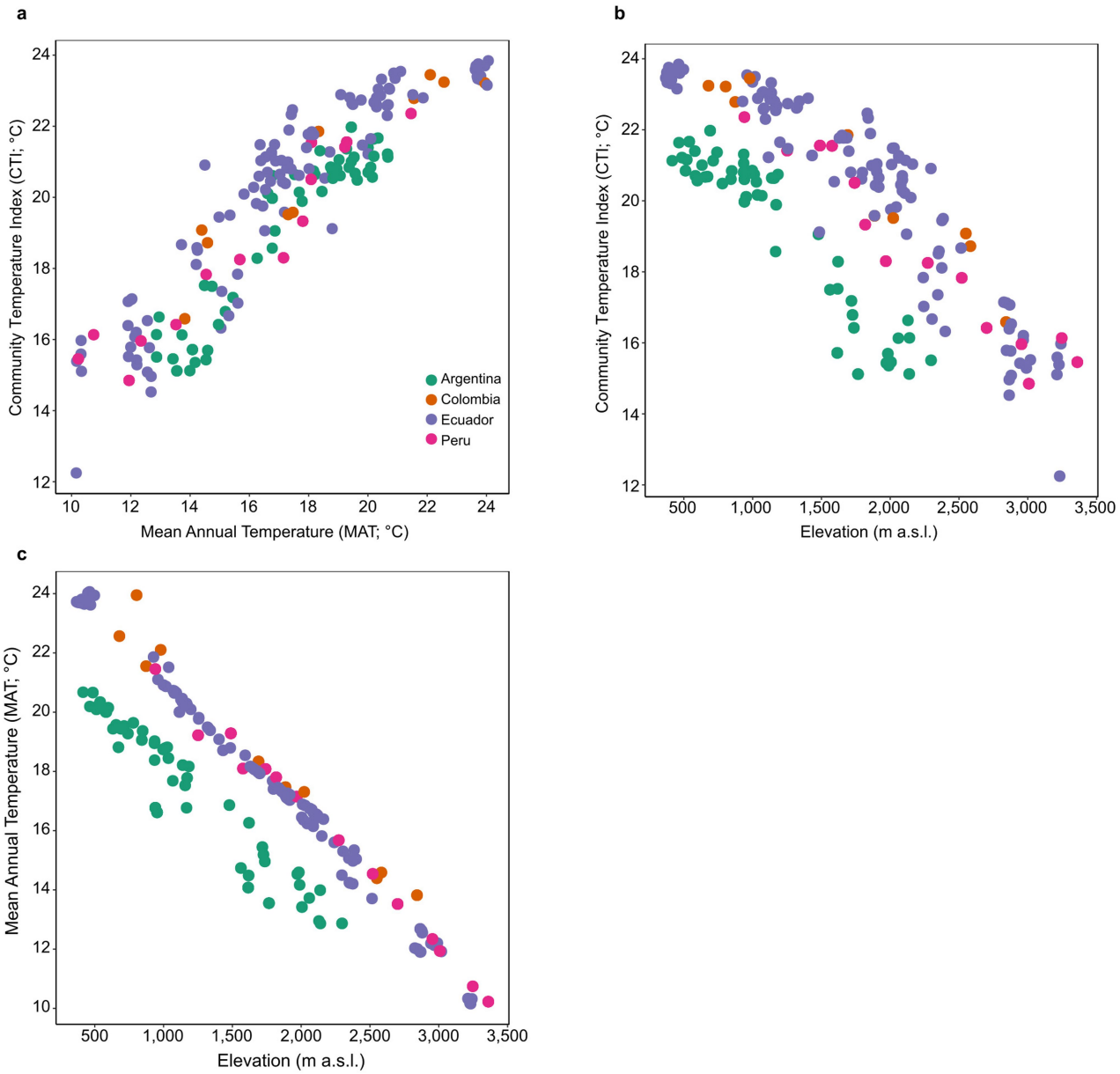
Species richness. For 1-ha plots with multiple censuses, we calculated species richness as the count of species that were found in the focal plot. We combined all morpho-species according to their genus assignments and added the genus to the species counts. We performed a linear model between MAT and species richness.
Range of thermal optima within plots. For plots with multiple censuses, we calculated the range of thermal optima (SpOptT) for all species that occurred within each plot as the difference between maximum and minimum SpOptT of the co-occurring species. We performed a linear model between the range of thermal optima and TR_{plot} values of the plots.
Change in basal area of more- versus less-thermophilic species per plot. For plots with multiple censuses, we calculated the change in basal area per plot generated by recruitment and growth (increase in basal area) versus mortality (decrease in basal area) for more-thermophilic species (that is, species with thermal optima above the CTI of a plot) and less-thermophilic species (that is, species with thermal optima below the CTI of a plot). For this analysis, we only included species for which the species thermal optima (SpOptT) were calculated using species-level GBIF records; we did not include species for which the thermal optima were estimated based on the distribution of congeners. We standardized the change in basal area by plot size and express the change as a percentage of the initial basal area. We performed a loess regression analysis between the basal area of more- versus less-thermophilic species and the MAT of the plots.
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The plot data that support the findings of this study are available from the Red de Bosques (https://redbosques.condesan.org/) upon reasonable request. The list of species included in the analysis with their number of GBIF records after filtering and their estimated thermal optima is available in Supplementary Table 2.

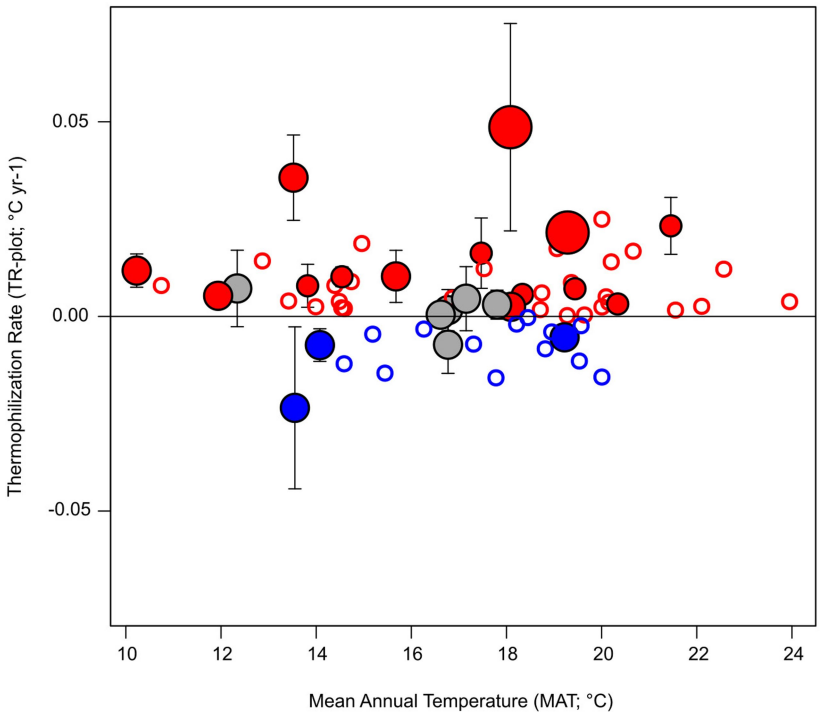
45. Bivand, R. & Yu, D. spgwr: geographically weighted regression. R package version 0.6-32 https://CRAN.R-project.org/package=spgwr (2017).
46. Dobrowski, S. Z. A climatic basis for microrefugia: the influence of terrain on climate. *Glob. Change Biol.* **17**, 1022–1035 (2011).
47. Hijmans, R. J. & van Etten, J. raster: geographic analysis and modeling with raster data. R package version 2.0-12 http://CRAN.R-project.org/package=raster (2012).
48. The Plant List. The plant list: a working list of all plant species. version 1.1 http://www.theplantlist.org/ (2013).
49. Flann, C. Global Compositae Checklist. https://doi.org/10.15468/g7yhgt (2009).
50. White, R. ILDIS: The International Legume Database and Information Service. http://www.ildis.org/ (accessed 13 May 2016).
51. Missouri Botanical Garden. Tropicos. http://www.tropicos.org (accessed 13 May 2016).
52. USDA NRCS. The PLANTS Database. http://plants.usda.gov (accessed 13 May 2016).
53. Feeley, K. J. & Silman, M. R. Biotic attrition from tropical forests correcting for truncated temperature niches. *Glob. Change Biol.* **16**, 1830–1836 (2010).

© 2018 Springer Nature Limited. All rights reserved.



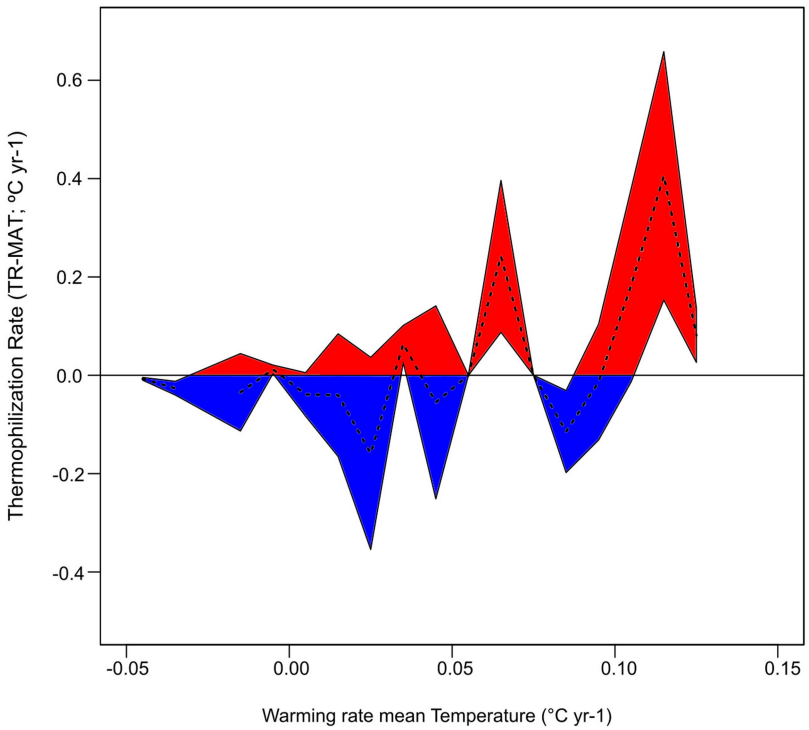
Extended Data Fig. 1 | CTI, MAT and elevation of study plots. a, The relationship between the mean CTI for each of the Andean forest plots (averaged across all censuses) and the MAT at the plot locations. $n = 186$, slope = 0.71, $R = 0.92$, 95% confidence interval = 0.88–0.93, $P < 0.001$. **b**, The relationship between the mean plot CTI and plot elevation. $n = 186$, $R = -0.77$, 95% confidence interval = −0.82 to −0.7, $P < 0.001$. **c**, The relationship between plot MAT and plot elevation. $n = 186$, $R = -0.92$, 95% confidence interval = −0.93 to −0.88, $P < 0.001$. All analyses are two-sided Spearman correlations.

© 2018 Springer Nature Limited. All rights reserved.



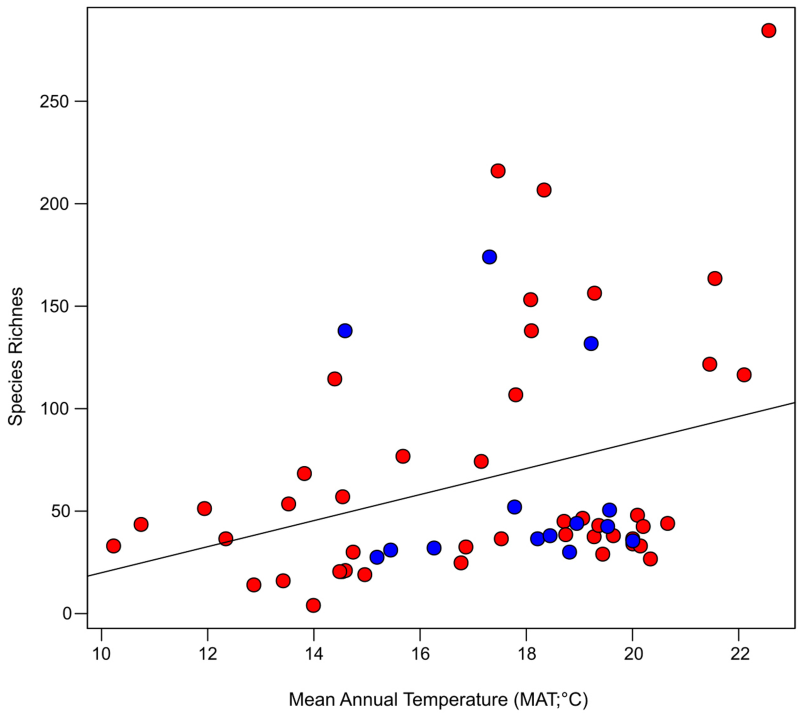
Extended Data Fig. 2 | Regression-based thermophilization rates of repeatedly censused plots. TR_{plot} was compared to the MAT for the Andean forest plots with multiple censuses ($n = 64$). Each point represents one plot and the size of the point is proportional to the number of censuses. Error bars are 95% confidence intervals based on the linear least-square regressions of the CTI versus census year of each plot. Grey points represent plots with non-significant TR_{plot} values and filled, coloured points represent plots with significant TR_{plot} values; hollow points are plots with only two censuses and for which the significance of the TR_{plot} could therefore not be determined. Positive and negative TR_{plot} are coloured red and blue, respectively.

© 2018 Springer Nature Limited. All rights reserved.



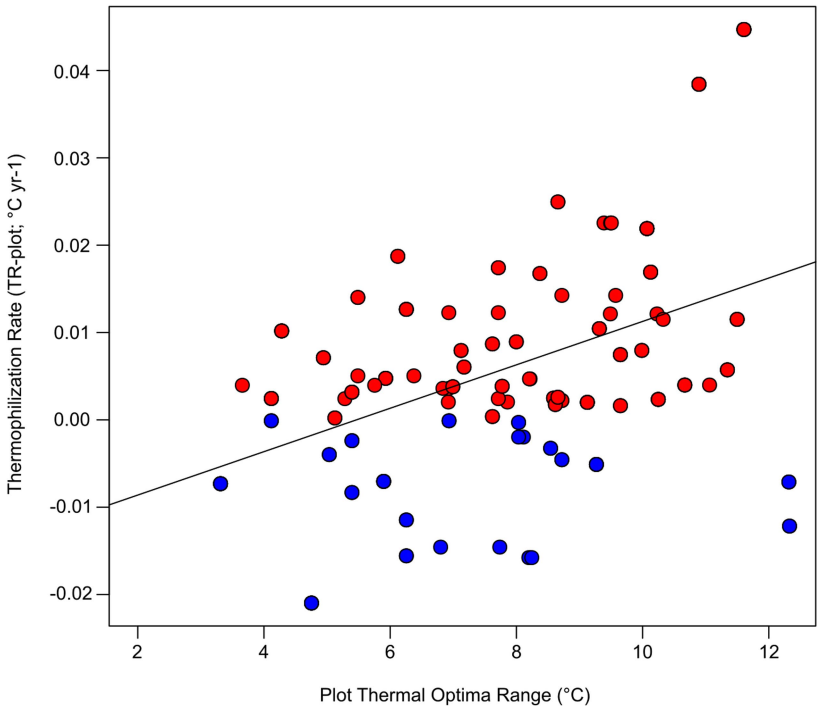
Extended Data Fig. 3 | Thermophilization rates for areas with different warming rates. The thermophilization rates in areas with different warming rates (TR_{warm}; the annualized change in the mean CTI of all plots within a band of equitable warming rate) were compared to the warming rate. $n = 283$ plot censuses, assigned to 20 warming bands. The dashed line indicates the mean TR_{warm} and the coloured shaded area indicates the 95% confidence interval of TR_{warm}. Positive and negative TR_{warm} is coloured red and blue, respectively.

© 2018 Springer Nature Limited. All rights reserved.



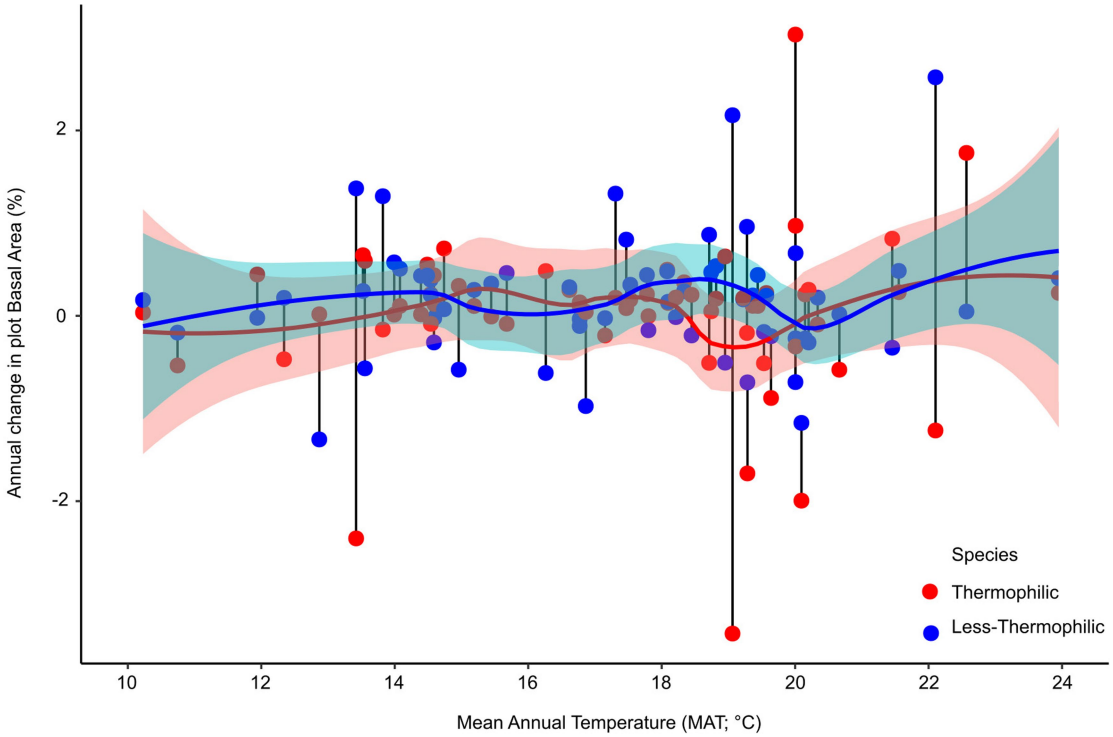
Extended Data Fig. 4 | Species richness of repeatedly censused 1-ha plots. Species richness versus MAT in the 1-ha Andean forest plots with multiple censuses. $n = 61$. Each point represents one plot and the red and blue colours indicate positive and negative TR_{plot} values, respectively. The line shows the linear regression between MAT and species richness. $R^2 = 0.10$, $P < 0.05$.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Range of the thermal optima in study plots. The range of thermal optima of co-occurring species versus TR_{plot} in the plots with multiple censuses. $n = 64$. Each point represents one plot and the red and blue colours represent positive and negative TR_{plot} values, respectively. The line shows the linear regression between the range of the thermal optima of the plots and TR_{plot}. $R^2 = 0.19$, $P < 0.001$.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Basal area change in species composition. Percentage change in absolute basal area per plot for more-thermophilic (species thermal optimum > plot CTI) and less-thermophilic (species thermal optimum < plot CTI) species versus MAT in plots with multiple censuses ($n = 64$). The more- and less-thermophilic species are coloured red and blue, respectively. Lines show loess regression fits between the percentage change in basal area and MAT, and the shaded areas represent the 95% confidence intervals around the loess regressions.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Results for alternative calculations of TR_{MAT} and TR_{plot}

		All data	Stems ≥ 10cm dbh	Plots size ≥ 1ha
	N. plots	186	186	79
	N. plots with ≥ 2 census	64	64	61
TR_{MAT}	CTI ~ Year +(1|plot id)	0.0029	0.0029	0.0061
	St. Error	0.001	0.001	0.001
TR_{plot}	pbinom	<0.001	<0.001	<0.001
	Upward m yr−1	1.7	1.7	2.1
	% Plots positive	72	72	75
	% Plots negative	28	28	25
	$TR_{interval}$ ~ +(1|plot id)	0.0066	0.0066	0.0084
	St. Error	0.002	0.002	0.002
	t.test	<0.001	<0.001	<0.001

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0772-0

A circuit from hippocampal CA2 to lateral septum disinhibits social aggression

Felix Leroy[1]*, Jung Park[1], Arun Asok[1], David H. Brann[1], Torcato Meira[1,2,3], Lara M. Boyle[1], Eric W. Buss[1], Eric R. Kandel[1,4] & Steven A. Siegelbaum[1]*

Although the hippocampus is known to be important for declarative memory, it is less clear how hippocampal output regulates motivated behaviours, such as social aggression. Here we report that pyramidal neurons in the CA2 region of the hippocampus, which are important for social memory, promote social aggression in mice. This action depends on output from CA2 to the lateral septum, which is selectively enhanced immediately before an attack. Activation of the lateral septum by CA2 recruits a circuit that disinhibits a subnucleus of the ventromedial hypothalamus that is known to trigger attack. The social hormone arginine vasopressin enhances social aggression by acting on arginine vasopressin 1b receptors on CA2 presynaptic terminals in the lateral septum to facilitate excitatory synaptic transmission. In this manner, release of arginine vasopressin in the lateral septum, driven by an animal's internal state, may serve as a modulatory control that determines whether CA2 activity leads to declarative memory of a social encounter and/or promotes motivated social aggression.

Considerable progress has been made in characterizing the neural circuits that underlie social aggression, a classic motivated behaviour[1,2]. However, less is known about how higher brain regions engaged in cognitive processing influence the decision to engage in aggression. Two subcortical regions that are important for aggression are the ventrolateral subnucleus of the ventromedial hypothalamus (VMHvl)[3,4] and the lateral septum (LS)[5,6], which contains exclusively GABAergic (γ-aminobutyric acid producing) inhibitory neurons[7]. Whereas excitation of VMHvl triggers aggression[3], inhibitory input to VMHvl from LS suppresses aggression[6]. As LS receives its most prominent input from the hippocampus[8], which has a critical role in declarative learning and memory, we investigated whether and how hippocampal output to LS may regulate aggressive behaviour. As many animals, including rodents and humans, form complex social hierarchies that influence aggressive behaviour[9,10], mnemonic information from the hippocampus about social identity could affect the decision to engage in aggression. Here we focused on the role of the relatively unexplored CA2 region of the hippocampus[11,12] in the control of social aggression. CA2 is of particular interest as it is both important for social memory[13,14] and highly enriched in the arginine vasopressin (AVP) 1b receptor (AVPR1b)[15], activation of which by the social neuropeptide AVP promotes aggression[16]. We report here that CA2 strongly promotes social aggression by acting through a CA2–LS–VMHvl disinhibitory circuit that is upregulated by AVP, providing an anatomical, functional and behavioural link between canonical circuits for memory and motivated behaviour.

CA2 projects to the dorsal LS

To gain insight into how CA2 may regulate behaviour we first examined its extra-hippocampal projections, focusing on dorsal CA2 (dCA2), which has been implicated in social memory[13,14]. We expressed channelrhodopsin2 labelled with enhanced yellow fluorescent protein (ChR2–eYFP) as an anterograde marker in dCA2 pyramidal neurons by injecting a Cre-dependent adeno-associated virus (AAV) in dCA2 of Amigo2-Cre mice, in which Cre expression is largely limited to CA2 pyramidal neurons[13] (Fig. 1a, b). We observed a dense network of CA2 fibres in the dorsal lateral septum (dLS) (Fig. 1c, Extended Data Fig. 1 and Supplementary Video 1), confirming conventional tracing[17]. Although CA3 pyramidal neurons also project to LS[8], fibres from mid-CA3 (CA3b) projected to the border of the ventricles and towards the ventral lateral septum (vLS), distinct from the site of CA2 projections, which target dLS closer to the midline (although there is some overlap in projections; Extended Data Fig. 2).

To investigate whether dCA2 forms synapses in dLS, we injected a retrograde tracer, cholera toxin β-subunit conjugated to Alexa 488 (CTB-488), into dLS (Extended Data Fig. 3a). Two weeks after injection, we observed strong labelling of dCA2 neurons co-labelled with the CA2 pyramidal neuron marker PCP4 (Extended Data Fig. 3b–d). As CTB can travel retrogradely across more than one synapse, we verified the monosynaptic nature of the projection by injecting G-deleted rabies virus expressing green fluorescent protein (GFP) into dLS (Fig. 1d). This labelled local dLS neurons near the injection site, consistent with local recurrent inhibition[18,19] (Extended Data Fig. 3e–g). In the hippocampus, retrogradely labelled cells were found in CA3, CA2, CA1 and in the fasciola cinerea, a region in the dorsal hippocampus that contains molecularly defined CA2 pyramidal neurons[20] (Fig. 1e–i). Both CA2 and fasciola cinerea retrogradely labelled neurons were co-labelled with antibodies against PCP4 and RGS14 (Fig. 1e–i).

As CA2 also sends a strong projection to dorsal CA1 (dCA1)[13,21], we sought to determine the proportion of CA2 pyramidal neurons that project to both dCA1 and dLS[22]. We injected CTB-488 into dLS and a monosynaptic retrograde Cre-dependent herpes simplex virus (HSV) expressing mCherry into dCA1 of Amigo2-Cre mice (Extended Data Fig. 4a). We found that 55 ± 6% (mean ± s.e.m.) of dCA2 pyramidal neurons were labelled with CTB-488, 44 ± 6% with mCherry, and 21 ± 4% with both markers (Extended Data Fig. 4b–d). This fraction of doubly labelled cells was identical to that expected if a single population of randomly labelled CA2 pyramidal neurons projected both to CA1 and LS (Extended Data Fig. 4d), suggesting that most CA2 pyramidal neurons project to both areas.

[1]Department of Neuroscience, The Kavli Institute for Brain Science, Mortimer B. Zuckerman Mind Brain Behavior Institute, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA. [2]Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal. [3]ICVS/3B's, PT Government Associate Laboratory, Braga/Guimarães, Portugal. [4]Howard Hughes Medical Institute, Columbia University, New York, NY, USA. *e-mail: felxfel@aol.com; sas8@columbia.edu

© 2018 Springer Nature Limited. All rights reserved.

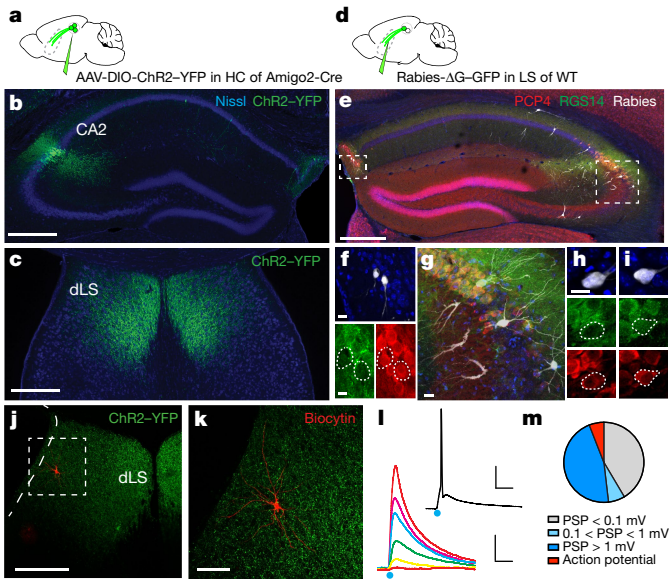


Fig. 1 | CA2 projections to the lateral septum. a, Anterograde tracing from dCA2 pyramidal neurons. HC, hippocampus. **b**, Coronal hippocampal section from an Amigo2-Cre mouse injected in dCA2 with rAAV5-EF1a-DIO-hChR2(E123T/T159C)–eYFP (green) and stained for Nissl (blue). **c**, Coronal section showing dCA2 projections in dLS. **d**, Monosynaptic retrograde tracing from dLS using G-deleted rabies virus expressing GFP (Rabies-ΔG–GFP). **e**, Coronal hippocampal section of a wild-type (WT) mouse injected in dLS with Rabies-ΔG–GFP (white), with immunostaining for CA2 markers RGS14 (green) and PCP4 (red). **f**–**i**, High-magnification views of genetically defined CA2 pyramidal neurons in fasciola cinerea (**f**) and CA2 proper (**h**, **i**) labelled using rabies virus (white) and co-labelled for RGS14 and PCP4. **j**, **k**, Coronal LS section from an Amigo2-Cre mouse injected into CA2 with rAAV5-EF1a-DIO-hChR2(E123T/T159C)–eYFP. A dLS cell filled with biocytin (red) during whole-cell recordings is shown at low (**j**) and high (**k**) magnifications. **l**, dLS cell current-clamp responses to photoactivation of ChR2-expressing dCA2 inputs with increasing light intensity. Inset shows spiking response of same dLS cell to light. **m**, Pie chart showing maximal dLS light-induced PSP amplitude frequency distribution, including fraction of cells in which an action potential was elicited (135 dLS neurons). Three mice were injected per experiment in **a**, **d**. All mice presented similar staining patterns. Scale bars: **b**, **c**, **e**, **j**, 400 μm; **f**–**i**, 20 μm; **k**, 100 μm; **l**, bottom, 5 mV/20 ms, top, 20 mV/20 ms.

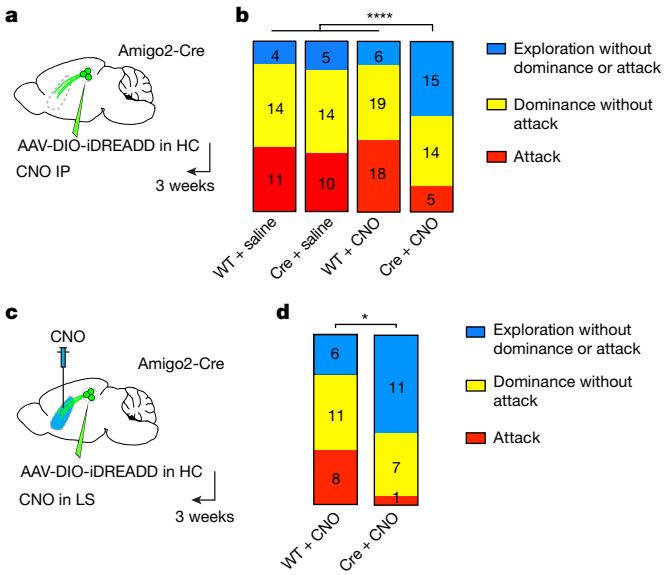


Fig. 2 | Silencing of CA2 or the CA2–LS projection inhibits social aggression. a, Amigo2-Cre mice (Cre) and wild-type (WT) littermates were injected into dCA2 with rAAV2-hsyn-DIO-HA-hM4D(Gi)-IRES-mCitrine (AAV-DIO-iDREADD). After 3 weeks, mice were injected intraperitoneally with CNO or saline and, 30 min later, subjected to the resident–intruder test. **b**, The numbers and proportions of mice exhibiting only social exploration (blue), exploration followed by social dominance (yellow) or exploration followed by dominance followed by at least one biting attack (attack, red). χ^2 tests, $^{****}P < 0.0001$. **c**, Silencing of dCA2 projections to LS. AAV-DIO-iDREADD was injected into dCA2 of wild-type and Amigo2-Cre mice and CNO was infused into dLS 20 min before testing. **d**, Stacked bar charts colour-coded as in **b**. χ^2 test, $^{*}P = 0.028$.

Dorsal CA2 strongly excites its dLS targets

To determine the synaptic influence of dCA2 on dLS, we expressed ChR2–eYFP in dCA2 pyramidal neurons and measured the electrophysiological responses of both CA2 pyramidal neurons and dLS neurons using whole-cell recordings. Most CA2 pyramidal neurons reliably fired action potentials in response to single or multiple 1-ms light pulses (Extended Data Fig. 5a–d). CA2 activation by a single light pulse elicited a large depolarizing postsynaptic potential (PSP) in around 50% of dLS cells (PSP peak = 6.9 ± 0.9 mV; Fig. 1j–m and Extended Data Fig. 5e, f) with a short latency (2.1 ± 0.1 ms; Extended Data Fig. 5g). As the PSP is the sum of CA2-evoked synaptic excitation and synaptic inhibition, we isolated the excitatory postsynaptic potential (EPSP) in dLS by applying $GABA_A$ and $GABA_B$ receptor antagonists SR 95531 and CGP 55845. Blockade of inhibition increased the peak amplitude of the PSP by 58 ± 14%, indicating the presence of an underlying inhibitory postsynaptic potential (IPSP; Extended Data Fig. 5h). However, even with inhibition intact, a single maximal-intensity light pulse elicited an action potential in about 12% of dLS cells (Fig. 1m). Subtraction of the EPSP from the net PSP at maximal intensity stimulation yielded an inferred IPSP of −3.0 ± 0.8 mV. Both the EPSP and the IPSP were blocked by the NMDA and AMPA glutamate receptor antagonists D-2-amino-5-phosphonovalerate (AP5) and 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) (PSP decreased to 11 ± 4% of baseline; Extended Data Fig. 5i), indicating that the IPSP resulted from disynaptic inhibition.

Silencing CA2–LS synapses inhibits social aggression

To investigate the role of dCA2 and its output to LS in aggression, we injected Cre-dependent AAV into dCA2 of Amigo2-Cre mice (and wild-type littermate controls) to express the inhibitory G-protein coupled receptor hM4Di (iDREADD[23]; Fig. 2a). After 3 weeks, we performed a resident–intruder test of aggression[3,24] by exposing a singly housed male mouse to a BALB/cJ male intruder. Both wild-type and iDREADD-expressing mice were injected with either saline or the iDREADD agonist clozapine-*N*-oxide (CNO, 10 mg kg^{-1}) intraperitoneally 30 min before testing.

Aggression in the resident–intruder test is characterized by a series of escalating behaviours over time, progressing from non-aggressive social exploration of the intruder (anogenital and facial sniffing), to social dominance (excessive grooming, chasing and/or mounting of the intruder) and then to one or more biting attacks, often preceded by tail rattling[24–26] (Supplementary Video 2). The behaviour of each resident mouse was categorized by its maximal level of aggressive behaviour (social exploration, social dominance or attack) during the 10-min test period.

All residents showed an initial period of social exploration, with the majority progressing to social dominance. In three control groups—wild-type mice injected with iDREADD AAV and given CNO (WT + CNO) or saline (WT + saline), and Amigo2-Cre mice injected with iDREADD AAV and given saline (Cre + saline)—roughly half of the residents escalated their behaviour further by engaging in one or more biting attacks. WT + saline and Cre + saline control mice showed the same incidence of aggression (31% compared to 35% of mice engaged in attack, respectively; χ^2 test, $P = 0.7$; Fig. 2b), demonstrating that expression of Cre or iDREADD did not alter aggression in the absence of CNO. Furthermore, the WT + saline and WT + CNO groups also showed similar levels of aggression, showing that CNO

© 2018 Springer Nature Limited. All rights reserved.

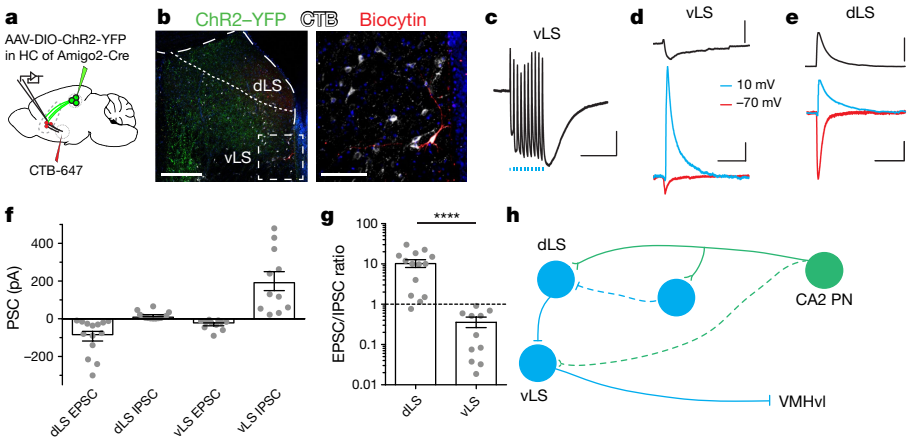


Fig. 3 | dLS neurons are excited by dCA2 projections and inhibit vLS cells. a, Amigo2-Cre mice were injected into dCA2 with rAAV5-EF1a-DIO-hChR2(E123T/T159C)–eYFP (AAV-DIO-ChR2–eYFP); a subset of mice was also injected into VMHvl with CTB-647. **b**, Left, coronal LS slice from mouse expressing ChR2–eYFP (green) in dCA2 projections. A vLS cell was filled with biocytin (red) during patch-clamp recording. Right, vLS neurons labelled with retrograde tracer CTB-647 (white) injected into VMHvl. **c**, Current-clamp voltage response of a vLS neuron to photostimulation of ChR2-expressing dCA2 inputs using a 10-pulse train of 1-ms light pulses at 20 Hz. **d**, **e**, Synaptic voltage response (top) and EPSC (red) and IPSC (blue) responses (bottom) of a vLS (**d**) or dLS (**e**) cell to a single 1-ms light pulse. **f**, Photoactivated EPSC and IPSC peak amplitudes from dLS and vLS cells. Grey dots, individual cells (14, 14, 11, 11 cells from left to right). Mean ± s.e.m. **g**, EPSC/IPSC ratios for dLS and vLS cells (14 and 11 cells from 10 mice). Mean ± s.e.m. Two-sided Mann–Whitney test, $****P < 0.0001$. **h**, Schematic of proposed CA2→dLS→vLS→VMHvl circuit. Blue, inhibitory LS neurons and synapses. Scale bars: **b**, left, 400 μm, right, 100 μm; **c**, 2 mV/500 ms; **d**, **e**, top, 5 mV; bottom, 40 pA/100 ms.

alone had no effect (31% versus 41%; χ^2 test, $P = 0.6$). However, silencing of CA2 (Cre + CNO group) caused a marked decrease in aggression compared to controls, with a smaller percentage of animals engaging in attack (15% compared to 36%) and a larger percentage showing only social exploration (44% compared to 15%; Fig. 2b). CA2 silencing also decreased the number of bites, attacks, tail rattles and total attack duration (Extended Data Fig. 6). The decrease in aggression did not result from general behavioural inhibition, as silencing of CA2 had no effect on locomotion, anxiety, object exploration or sociability[13] (Extended Data Fig. 7a–f). Finally, CA2 silencing had no effect on predator–prey aggression, indicating that CA2 was selectively required for social aggression (Extended Data Fig. 7g, h).

To determine whether dCA2 promotes aggression through its projections to LS, we expressed iDREADD in dCA2 and used a cannula to deliver CNO to LS (Fig. 2c). Application of CNO (5 μM) to septal slices from mice expressing iDREADD and ChR2 in dCA2 pyramidal neurons decreased the light-evoked PSP in dLS neurons to 37 ± 7% of baseline (Extended Data Fig. 8a), confirming the efficacy of terminal silencing. We examined the behavioural effect of infusing 1 μl of 1 mM CNO into dLS 20 min before the resident–intruder test. As shown in Fig. 2d, this significantly decreased the fraction of mice that engaged in attack (5% of Cre + CNO mice versus 32% of WT + CNO mice) and increased the fraction that showed only social exploration (58% Cre + CNO versus 24% WT + CNO, χ^2 test, $P < 0.0001$). The proper delivery of CNO into dLS was verified by infusion of the dye miniRuby through the cannula (Extended Data Fig. 8b). CNO infusion in LS is unlikely to act by diffusing to dCA2 because CNO infusion into the ventral hippocampus (vHPC), which is closer to dCA2 than is dLS, does not decrease aggression[23]. Thus, we conclude that dCA2 promotes aggression, at least in part, through its projections to dLS.

dCA2 inhibits vLS neurons that project to VMHvl

How can our result that dCA2 promotes aggression be reconciled with the classic action of LS to inhibit aggression[5,6], given that CA2 excites LS? One clue came from our finding that dCA2 projects largely to dLS, whereas the projections to VMHvl that inhibit aggression come primarily from vLS[6,27]. As anatomical results suggest that dLS neurons send inhibitory projections to vLS[18,28], we surmised that excitation of dLS neurons by dCA2 may inhibit vLS neurons, including those that tonically inhibit VMHvl, thereby forming a trisynaptic disinhibitory circuit that promotes aggression (Fig. 3h).

To determine whether dCA2 evokes disynaptic inhibition of vLS, we expressed ChR2–YFP in dCA2 pyramidal neurons and obtained whole-cell recordings from vLS cells in LS slices (Fig. 3a, b). In contrast to the large depolarizing response recorded in dLS (Fig. 1l), photostimulation of dCA2 inputs produced a large hyperpolarization in most vLS neurons (Fig. 3c). Voltage-clamp recordings showed that photostimulation evoked only a weak excitatory postsynaptic current (EPSC; measured with the membrane held at −70 mV) in vLS that was much smaller than the EPSC evoked in dLS (Fig. 3d–f). By contrast, photostimulation evoked a much larger inhibitory postsynaptic current (IPSC; measured at +10 mV) in vLS than in dLS (Fig. 3d–f). As a consequence, the EPSC/IPSC ratio was more than 30-fold larger in dLS than in vLS (dLS, 10.9 ± 2.4 versus vLS, 0.33 ± 0.1; Fig. 3g). Both dLS and vLS IPSCs resulted from feedforward inhibition in response to CA2 activation, as the latency of the EPSC was shorter than that of the IPSC (Fig. 3d, e) and the IPSC was almost completely suppressed by CNQX and AP5 (vLS IPSC decreased to 5.4 ± 1.9%; Extended Data Fig. 5j, k).

According to the disinhibition hypothesis, activation of dCA2 should produce disynaptic inhibition in the subset of vLS neurons that sends output to VMHvl (Fig. 3h). To test this idea, we identified vLS projection neurons by injecting the retrograde tracer CTB into VMHvl, which confirmed that vLS forms synapses with this target[6,27] (Fig. 3a, b). Whole-cell recordings from visually identified CTB^+ vLS cells revealed that their synaptic response to photostimulation of dCA2 inputs was indeed dominated by inhibition (Fig. 3c). We verified a polysynaptic connection from CA2 to VMHvl by injecting into VMH a trans-synaptic replication-competent HSV expressing mCherry that propagates retrogradely through several synaptic contacts[29]; this protocol labelled VMHvl, vLS, dLS and dCA2 (Extended Data Fig. 9a–c).

Finally, we investigated whether output from dCA2 enhances VMHvl activity during aggressive social behaviour by measuring the effect of silencing dCA2 on the number of c-Fos-labelled cells in VMHvl, a marker of neuronal activity known to be increased by aggression[30]. After confirming that aggression increased c-Fos labelling in our behavioural paradigm (Extended Data Fig. 10b), we tested whether silencing CA2 affected c-Fos levels. Injection of CNO caused a twofold decrease in the number of $c\text{-}Fos^+$ cells in the VMHvl of Cre^+ mice expressing iDREADD compared to wild-type mice injected with CNO

© 2018 Springer Nature Limited. All rights reserved.

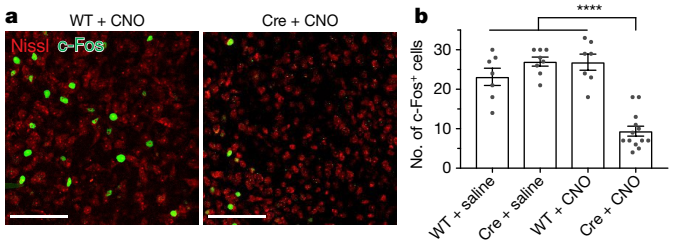


Fig. 4 | Silencing CA2 with iDREADD expression and CNO injection decreases c-Fos expression in VMHvl. a, Representative images of c-Fos immunofluorescence in VMHvl following bouts of aggression 30 min after intraperitoneal injection of CNO or saline in wild-type or Amigo2-Cre mice (all groups previously injected into CA2 with rAAV2-hsyn-DIO-HA-hM4D(Gi)-IRES-mCitrine). Scale bars, 100 μm. **b**, Number of c-Fos$^+$ cells in the three control groups (WT + saline, Cre + saline, WT + CNO) and the experimental group (Cre + CNO). Grey dots, individual mice (7, 8, 7 and 13 mice from left to right). Mean ± s.e.m. Two-way ANOVA genotype × drug interaction: $F = 42.9$, $^{****}P < 0.0001$.

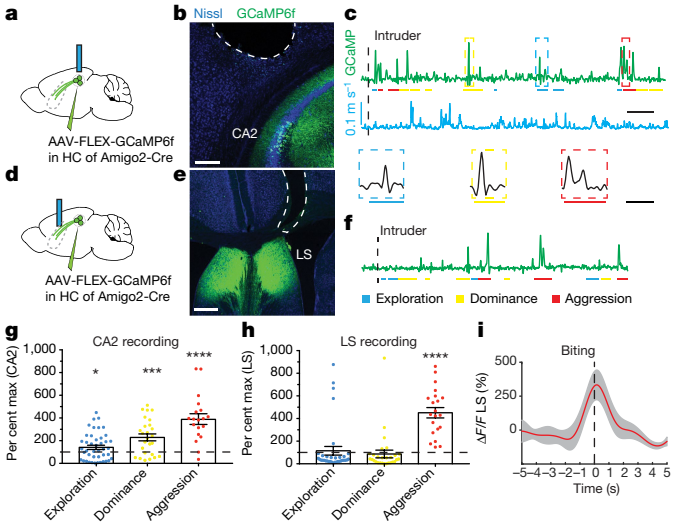


Fig. 5 | CA2 pyramidal neurons respond to diverse social encounters whereas CA2 terminals in LS selectively convey aggression-related information. a, Schematic of fibre photometry recordings of GCaMP6f fluorescence signals from CA2 pyramidal neurons in the hippocampus (four mice). **b**, Coronal section of the hippocampus showing the expression of GCaMP6f and optical fibre location (dashed outline). **c**, Example recording of GCaMP6f signal (green) with fibre over CA2 during resident–intruder test. Mouse velocity shown in blue. Coloured bars define episodes of social exploration (blue), social dominance (yellow) and biting attack (red). Coloured dashed rectangles represent magnified examples of GCaMP6f signal during each type of interaction. **d**, Schematic of GCaMP6f recordings with fibre over dLS (five mice). **e**, Coronal section of LS showing expression of GCaMP6f and fibre location. **f**, Example recording of GCaMP6f signals from CA2 pyramidal neuron projections in dLS during resident–intruder test. Coloured bars as in **c**. **g**, Plot of CA2 peak GCaMP6f signals during behaviours. Each point is from a single behavioural episode (43, 26 and 19 episodes left to right, five mice). Mean ± s.e.m. One-sample two-sided *t*-tests against baseline (from left to right): $^*P = 0.02$, $^{***}P = 0.0002$, $^{****}P < 0.0001$. **h**, As in **g** but recorded from dLS (34, 28 and 21 episodes from left to right, 6 mice), $^{****}P < 0.0001$. **i**, Mean GCaMP6f signal (red line, mean; grey area, ± s.e.m.) in dLS aligned to onset of biting and normalized to baseline (21 attack episodes). Scale bars: **b**, 200 μm; **e**, 400 μm; **c**, **f**, 50 s; enlarged episodes in **c**, 10 s.

or the two other control groups (Fig. 4, Extended Data Fig. 10a). The decrease in c-Fos labelling upon CA2 silencing was not secondary to the decreased fraction of mice showing aggression, as we restricted the c-Fos analysis to the subset of mice that exhibited one or more biting attacks in control and experimental groups (Extended Data Fig. 10c, see Methods). Thus, we conclude that CA2 output normally enhances VMHvl activity during aggression, presumably by activating the LS–VMHvl disinhibitory circuit.

CA2 output to LS increases during social aggression

To assess whether CA2 is activated during social exploration and aggression, we injected a Cre-dependent AAV into dCA2 of Amigo2-Cre mice to express the genetically encoded fluorescent Ca^{2+} sensor GCaMP6f. We then used fibre photometry[31] to measure intracellular Ca^{2+} levels based on GCaMP6f fluorescence with a fibre over dCA2 (Fig. 5a–c). We saw little change in GCaMP6f signal in CA2 during non-social exploration or as a function of mouse velocity (Fig. 5c, Extended Data Fig. 11a). However, episodes of social exploration elicited a small but significant increase in GCaMP6f peak fluorescence to 147 ± 19% of baseline (Fig. 5c, g; one-sample two-sided *t*-test against baseline, $P = 0.02$). Episodes of social dominance were associated with a greater increase in Ca^{2+} to 234 ± 31% of baseline (Fig. 5c, g), significantly larger than during social exploration (two-sided *t*-test, $P = 0.01$). Ca^{2+} levels increased even further during aggression (biting attack) to 394 ± 47% of baseline (Fig. 5c, g), significantly greater than during social dominance (two-sided *t*-test, $P = 0.005$). Analysis of the mean GCaMP6f signal (rather than the peak) during the behavioural episodes yielded similar results (Extended Data Fig. 11b).

Next, we investigated whether CA2 input to LS was also regulated during social interactions by expressing GCaMP6f in CA2 pyramidal neurons with the fibre positioned over dLS (Fig. 5d–f). We observed a large increase in the peak GCaMP6f signal during aggression (456 ± 46%; Fig. 5f, h), similar to that seen in dCA2. However, unlike in dCA2, the increase in Ca^{2+} in dCA2 projections to dLS was highly selective for social aggression, with no significant change during social exploration (one-sample *t*-test, $P = 0.6$) or social dominance (one-sample *t*-test, $P = 0.8$). Analysis of mean GCaMP6f signals yielded similar results (Extended Data Fig. 11c). The increase in Ca^{2+} preceded bite onset by 1–2 s (Fig. 5i), suggesting that output from dCA2 to dLS contributes to attack.

In contrast to the large increase in dCA2 activity during social aggression, there was no significant change in Ca^{2+} levels during exploration of a novel object or environment or during feeding. (Extended Data Fig. 11d–j). There was a small but significant increase in dCA2 GCaMP6f signal during exploration of a novel female (147 ± 24%, two-sided one-sample *t*-test, $P = 0.03$) and during predator–prey aggression (150 ± 24%, two-sided one-sample *t*-test, $P = 0.047$), similar to the response during exploration of a male. However, both female exploration and prey aggression responses were significantly less than those during social aggression (two-sided *t*-tests, $P < 0.0001$; Extended Data Fig. 11i, j). The Ca^{2+} responses to a novel female tended to decline with repeated exposure and increased when a new animal was introduced (Extended Data Fig. 11k–m), suggesting that CA2 may encode social novelty, a factor known to promote aggression[32,33].

CA2 AVPR1b enhances LS input and aggression

As CA2 activity is required both for social memory acquired during non-aggressive social exploration and for social aggression, downstream circuits must differentiate when CA2 output should trigger aggression. Because aggression is regulated by an animal's internal state[1], we hypothesized that the social neuropeptide AVP might provide a state-dependent modulatory signal to enhance the ability of CA2 activity to trigger attack.

To test this idea, we first investigated whether AVP altered synaptic transmission between dCA2 and dLS by expressing ChR2 in dCA2 pyramidal neurons and recording light-evoked PSPs in dLS neurons (Fig. 6a). Bath application of 100 nM AVP increased the peak PSP to 182 ± 15% of its initial value (Fig. 6b), accompanied by a decrease in the PSP paired-pulse ratio to 87 ± 2% of its initial value (Extended Data Fig. 12a–c). This result indicates that AVP acts presynaptically to enhance transmitter release from dCA2 inputs.

© 2018 Springer Nature Limited. All rights reserved.

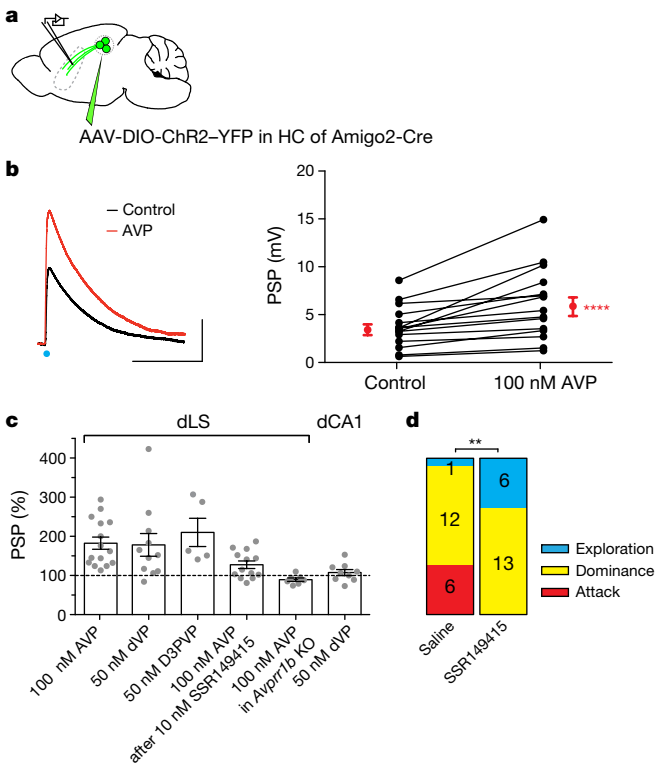


Fig. 6 | AVPR1b activation on CA2 presynaptic terminals in LS potentiates synaptic excitation and increases social aggression. a, AAV injection for ChR2 expression in dCA2. **b**, Current-clamp responses (PSPs) of dLS neurons following photostimulation of ChR2-expressing dCA2 inputs before (control) and 15 min after bath application of 100 nM AVP (15 cells). Scale bars, 2 mV and 100 ms. Black dots, individual cells; red dots, mean ± s.e.m. Two-sided Wilcoxon test, $^{****}P < 0.0001$. **c**, PSP amplitudes in LS neurons as per cent of initial PSP after application of (left to right): 100 nM AVP, 50 nM dVP, 50 nM D3PVP, 100 nM AVP in presence of 10 nM SSR149415, 100 nM AVP in an *Avpr1b* knockout (KO) mouse (15, 11, 5, 13, 7 cells). Bar on right, effect of 50 nM dVP on CA1 pyramidal neuron PSP evoked by photostimulation of CA2 inputs (9 cells). Mean ± s.e.m. **d**, Stacked bar charts representing fractions of mice exhibiting only social exploration (blue), exploration followed by dominance (yellow) or exploration, dominance and biting attack (red). Mice were infused with saline or SSR149415 into dLS before resident–intruder test. 19 mice per group. χ^2 test, $^{**}P = 0.008$.

As CA2 pyramidal neurons express AVPR1b[15] but not AVPR1a[34], which is expressed in LS neurons[19], we tested the effects of two AVPR1b-selective agonists (50 nM [dLeu4,Lys8]-AVP (dVP) or 50 nM [deamino-Cys1, D-3-pyridyl-Ala2, Arg8]-vasopressin (D3PVP)). Both compounds potentiated the PSP to the same extent as AVP (178 ± 29% for dVP and 210 ± 36% for D3PVP; Fig. 6c). Furthermore, the effect of AVP was reduced by the AVPR1b-specific antagonist SSR149415[35] (PSP, 121 ± 10% of baseline; two-sided Mann–Whitney test compared to AVP alone, $P = 0.01$; Fig. 6c) and eliminated by genetic deletion of *Avpr1b* (PSP, 89 ± 4% of baseline; two-sided Mann–Whitney test compared to AVP application in wild type, $P < 0.0001$; Fig. 6c). Although AVP can activate oxytocin receptors, the oxytocin agonist TGOT (250 nM) did not alter the PSP (97 ± 3% of baseline; Extended Data Fig. 12c, d). Unexpectedly, AVP did not alter the PSP recorded in dCA1 pyramidal neurons in response to photoactivation of dCA2 inputs in hippocampal slices (PSP, 107 ± 8% of baseline; two-sided Mann–Whitney test compared to AVP response in dLS, $P = 0.0002$; Fig. 6c), suggesting that any AVPR1b expressed in dCA2 terminals in dCA1 cannot regulate transmitter release.

Finally, we tested the behavioural importance of AVPR1b on dCA2 terminals in dLS by infusing either saline or SSR149415 through a cannula into dLS (Fig. 6d). The AVPR1b antagonist decreased the fraction of mice displaying aggression (from 32% to 0%) and increased the fraction displaying only social exploration (from 5% to 32%), suggesting that AVPR1b in dCA2 terminals in dLS may act as a state-dependent regulator of social aggression[36].

Discussion

Together with previous results, our data indicate that dCA2 pyramidal neurons are required both for social memory[13,14,23,34] and to promote social aggression[16]. The mnemonic function of dCA2 is mediated by its projections to ventral CA1[23] (Extended Data Fig. 1c), another hippocampal region that has been implicated in social memory[37]. Here we show that dCA2 promotes aggression through its output to dLS, which activates a circuit that disinhibits the VMHvl hypothalamic subnucleus, which has been implicated in aggression[1–4,6]. Thus, our findings provide a link, at both the behavioural and circuit levels, that connects the hippocampus, a brain region noted for its role in declarative memory, with the control of a motivated behaviour and its hypothalamic trigger.

How might mnemonic information provided by CA2 participate in regulating aggressive behaviour? It is likely that a decision to engage in social attack requires evaluation of memories of past social encounters that may predict the potential outcome of aggression. This decision is also likely to require a determination of social novelty, as aggression is triggered more readily by a novel than a familiar intruder[32,33].

Another question is why dCA2 social signals, which are generated during both non-aggressive and aggressive social encounters, trigger aggression only under certain circumstances. For example, aggression is observed routinely when a socially isolated male encounters a novel adult male, but rarely during encounters with a novel juvenile male or a novel female, both of which activate CA2 (Fig. 5, Extended Data Fig. 11). As most dCA2 neurons project to both CA1 and dLS, it is unlikely that separate CA2 subpopulations are activated for memory versus aggression. Rather, we suggest that the social signal conveyed by dCA2 to dLS may be modulated by the internal state of an animal through release of the social neuropeptide AVP to facilitate information transfer from dCA2 to dLS. As other hippocampal regions, notably dCA3, also project extensively to dLS[8], future studies will be needed to explore the relative roles of different hippocampal regions in social behaviour.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0772-0.

Received: 5 August 2017; Accepted: 12 October 2018;
Published online 5 December 2018.

1. Anderson, D. J. Circuit modules linking internal states and social behaviour in flies and mice. *Nat. Rev. Neurosci.* **17**, 692–704 (2016).
2. Hashikawa, K., Hashikawa, Y., Falkner, A. & Lin, D. The neural circuits of mating and fighting in male mice. *Curr. Opin. Neurobiol.* **38**, 27–37 (2016).
3. Lin, D. et al. Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470**, 221–226 (2011).
4. Falkner, A. L., Grosenick, L., Davidson, T. J., Deisseroth, K. & Lin, D. Hypothalamic control of male aggression-seeking behavior. *Nat. Neurosci.* **19**, 596–604 (2016).
5. Brady, J. V. & Nauta, W. J. H. Subcortical mechanisms in emotional behavior: affective changes following septal forebrain lesions in the albino rat. *J. Comp. Physiol. Psychol.* **46**, 339–346 (1953).
6. Wong, L. C. et al. Effective modulation of male aggression through lateral septum to medial hypothalamus projection. *Curr. Biol.* **26**, 593–604 (2016).
7. Risold, P. Y. & Swanson, L. W. Chemoarchitecture of the rat lateral septal nucleus. *Brain Res. Brain Res. Rev.* **24**, 91–113 (1997).
8. Risold, P. Y. & Swanson, L. W. Connections of the rat lateral septal complex. *Brain Res. Brain Res. Rev.* **24**, 115–195 (1997).
9. Zhou, T. et al. History of winning remodels thalamo-PFC circuit to reinforce social dominance. *Science* **357**, 162–168 (2017).
10. Williamson, C. M., Lee, W. & Curley, J. P. Temporal dynamics of social hierarchy formation and maintenance in male mice. *Anim. Behav.* **115**, 259–272 (2016).
11. Lorente De Nó, R. Studies on the structure of the cerebral cortex. II. Continuation of the study of the ammonic system. *J. Psychol. Neurol.* **46**, 113–177 (1934).

© 2018 Springer Nature Limited. All rights reserved.

12. Dudek, S. M., Alexander, G. M. & Farris, S. Rediscovering area CA2: unique properties and functions. *Nat. Rev. Neurosci.* **17**, 89–102 (2016).
13. Hitti, F. L. & Siegelbaum, S. A. The hippocampal CA2 region is essential for social memory. *Nature* **508**, 88–92 (2014).
14. Stevenson, E. L. & Caldwell, H. K. Lesions to the CA2 region of the hippocampus impair social memory in mice. *Eur. J. Neurosci.* **40**, 3294–3301 (2014).
15. Young, W. S., Li, J., Wersinger, S. R. & Palkovits, M. The vasopressin 1b receptor is prominent in the hippocampal area CA2 where it is unaffected by restraint stress or adrenalectomy. *Neuroscience* **143**, 1031–1039 (2006).
16. Pagani, J. H. et al. Role of the vasopressin 1b receptor in rodent aggressive behavior and synaptic plasticity in hippocampal area CA2. *Mol. Psychiatry* **20**, 490–499 (2015).
17. Cui, Z., Gerfen, C. R. & Young, W. S. III Hypothalamic and other connections with dorsal CA2 area of the mouse hippocampus. *J. Comp. Neurol.* **521**, 1844–1866 (2013).
18. Alonso, J. R. & Frotscher, M. Organization of the septal region in the rat brain: a Golgi/EM study of lateral septal neurons. *J. Comp. Neurol.* **286**, 472–487 (1989).
19. Allaman-Exertier, G., Reymond-Marron, I., Tribollet, E. & Raggenbass, M. Vasopressin modulates lateral septal network activity via two distinct electrophysiological mechanisms. *Eur. J. Neurosci.* **26**, 2633–2642 (2007).
20. Evans, P. R., Lee, S. E., Smith, Y. & Hepler, J. R. Postnatal developmental expression of regulator of G protein signaling 14 (RGS14) in the mouse brain. *J. Comp. Neurol.* **522**, 186–203 (2014).
21. Kohara, K. et al. Cell type-specific genetic and optogenetic tools reveal hippocampal CA2 circuits. *Nat. Neurosci.* **17**, 269–279 (2014).
22. Tamamaki, N., Abe, K. & Nojyo, Y. Three-dimensional analysis of the whole axonal arbors originating from single CA2 pyramidal neurons in the rat hippocampus with the aid of a computer graphic technique. *Brain Res.* **452**, 255–272 (1988).
23. Meira, T. et al. A hippocampal circuit linking dorsal CA2 to ventral CA1 critical for social memory dynamics. *Nat. Commun.* **9**, 4163 (2018).
24. Koolhaas, J. M. et al. The resident-intruder paradigm: a standardized test for aggression, violence and social stress. *J. Vis. Exp.* **4367**, e4367 (2013).
25. Blanchard, R. J. & Blanchard, D. C. Aggressive behavior in the rat. *Behav. Biol.* **21**, 197–224 (1977).
26. Takahashi, A., Quadros, I. M., de Almeida, R. M. M. & Miczek, K. A. in *Current Topics in Behavioral Neurosciences* **vol. 12** (eds Cryan, J. & Reif, A.) 73–138 (Springer, Berlin, 2011).
27. Toth, M., Fuzesi, T., Halasz, J., Tulogdi, A. & Haller, J. Neural inputs of the hypothalamic "aggression area" in the rat. *Behav. Brain Res.* **215**, 7–20 (2010).
28. Swanson, L. W. & Cowan, W. M. The connections of the septal region in the rat. *J. Comp. Neurol.* **186**, 621–655 (1979).
29. Ugolini, G. Advances in viral transneuronal tracing. *J. Neurosci. Methods* **194**, 2–20 (2010).
30. Lee, H. et al. Scalable control of mounting and attack by $Esr1^{+}$ neurons in the ventromedial hypothalamus. *Nature* **509**, 627–632 (2014).
31. Hashikawa, K. et al. $Esr1^{+}$ cells in the ventromedial hypothalamus control female aggression. *Nat. Neurosci.* **20**, 1580–1590 (2017).
32. Connor, J. L. & Lynds, P. G. Mouse aggression and the intruder-familiarity effect: evidence for multiple-factor determination c57bl. *J. Comp. Physiol. Psychol.* **91**, 270–280 (1977).
33. Szenczi, P., Bánszegi, O., Groó, Z. & Altbäcker, V. Development of the social behavior of two mice species with contrasting social systems. *Aggress. Behav.* **38**, 288–297 (2012).
34. Smith, A. S., Williams Avram, S. K., Cymerblit-Sabba, A., Song, J. & Young, W. S. Targeted activation of the hippocampal CA2 area strongly enhances social memory. *Mol. Psychiatry* **21**, 1137–1144 (2016).
35. Gal, C. S.-L. et al. An overview of SSR149415, a selective nonpeptide vasopressin V1b receptor antagonist for the treatment of stress-related disorders. *CNS Drug Rev.* **11**, 53–68 (2006).
36. Blanchard, R. J. et al. AVP V1b selective antagonist SSR149415 blocks aggressive behaviors in hamsters. *Pharmacol. Biochem. Behav.* **80**, 189–194 (2005).
37. Okuyama, T., Kitamura, T., Roy, D. S., Itohara, S. & Tonegawa, S. Ventral CA1 neurons store social memory. *Science* **353**, 1536–1541 (2016).

Acknowledgements We thank N. Renier, the Rockefeller imaging center and the laboratories of F. Polleux and T. Jessell for their help in creating Supplementary Video 1. We also thank R. Bruno, L. Herbaut and the members of the Siegelbaum laboratory for discussions. This work was supported by the R01 MH104602 and R01 MH106629 from NIH (S.A.S.), PD/BD/113700/2015 from the Portuguese Foundation for Science and Technology (T.M.), 5TL1TR001875-03 from NIH (E.W.B.) and the HHMI (E.R.K).

Reviewer information *Nature* thanks S. Dudek, C. Gross and D. Lin for their contribution to the peer review of this work.

Author contributions Conceptualization: F.L. and S.A.S. Viral injections: F.L., D.H.B. and T.M. Immunohistochemistry: F.L., D.H.B. and L.M.B. Intracellular recordings: F.L. Behavioral assays: F.L., J.P., T.M. and L.B. Fibrephotometry: F.L. and A.A. 3D representation: F.L. and E.W.B. Writing (original draft): F.L. Writing (review and editing): F.L., S.A.S. and E.R.K. Visualization: F.L. Supervision: S.A.S. and E.R.K. Funding acquisition: S.A.S. and E.R.K.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0772-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0772-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.L. or S.A.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Experimental models. All mouse procedures were performed in accordance with the regulations of the Columbia University IACUC. We used the following mouse lines: Amigo2-Cre mice and their wild-type littermates[13], and Grik4-Cre[38] and *Avpr1b*-K0 mice[39] crossed with Amigo2-Cre mice (heterozygous for Cre and homozygous for *Avpr1b*-K0), all on the C57Bl/6J background. For the social aggression tests, we used BALB/cJ intruders[3,16,39]. Tracing and in vitro recordings were performed on male and female mice. We observed no difference related to sex and the results were pooled together. Behavioural tests were performed on sexually naive male mice only. All mice were maintained on a 12-h light–dark cycle with ad libitum access to food and water. We used mice between 2 and 6 months old. No statistical methods were used to predetermine sample size, but sample sizes are consistent with those generally employed in the field. Animals were randomly assigned numbers and tested blind for the experimental conditions. All behavioural experiments were scored by an individual blind to the genotype and experimental design.

Viral injections. For all injections, mice were anaesthetized using isoflurane and given analgesics. A craniotomy was performed above the target region and a glass pipette was stereotaxically lowered to the desired depth. All coordinates are in millimetres with the Bregma as reference. Injections were performed using a nano-inject II (Drummond Scientific). Twenty-three nanolitres of solution was delivered every 15 s until the total amount was reached. The pipette was retracted after 5 min.

Hippocampal injections. We bilaterally injected 200 nl of the following viruses: rAAV5-EF1a-DIO-hChR2(E123T/T159C)-eYFP[40] (UNC, lot AV4828b), rAAVDJ-hSyn-FLEX-mGFP-2A-Synatophysin-mRuby (Sanford viral core, #GVVC-AAV-100, lot 1930), rAAV5-hsyn-DIO-eGFP (UNC, #4497), rAAV2-hsyn-DIO-HA-hM4D(Gi)-IRES-mCitrine[41] (Addgene, #50455 prepared by the Duke University vector core) and rAAV1-syn-FLEX-GCaMP6f-WPRE-SV40[42] (Addgene, #100833-AAV1) into the hippocampus of Amigo2-Cre or Grik4-Cre mice. Injection coordinates were the following: AP 2, ML ±1.8, DV –1.7. Incubation time was 3 weeks for immunohistochemistry or electrophysiological recording and 4 weeks for behaviour. Injection of rAAV5-EF1a-DIO-hChR2(E123T/T159C)-eYFP led to selective expression of ChR2–eYFP in 80 ± 3% (18 mice) of all CA2 pyramidal neurons in the dorsal half of the hippocampus.

Retrograde tracings from the LS. We bilaterally injected 200 nl of G-deleted rabies[43] SAD-B19-ΔG.mCherry (Salk Institute) or CTB conjugated to Alexa-488 (ThermoFisher Scientific, #C22841) into the dLS at the following coordinates: AP +0.3, ML ±0.1, DV –2.5.

Dual retrograde virus injection into LS and CA1. We injected 100 nl of the Cre-dependent retrograde monosynaptic HSV EF1a-LSIL-mCherry (MIT McGovern Institute vector core, cat# RN413) and 400 nl of CTB conjugated to Alexa-488 (ThermoFisher Scientific, #C22841) into dCA1 and dLS, respectively. dCA1 injection coordinates were the following: AP –2, ML ± 1.4, DV –1.7. dLS coordinates were as above. One week later, mice were perfused and processed for mCherry and RGS14 labelling.

Retrograde tracing from the VMHvl. We injected 400 nl of the trans-synaptic HSV CMV–mCherry (CNNV, #HSC373) or CTB-647 (ThermoFisher Scientific, #C34778) into the dLS at the following coordinates: AP –1.7, ML ± 0.68, DV 5.8.

Cannula guide implantation. Mice were implanted with a cannula guide extending 2.4 mm below the pedestal (Plastics One, #C315G 2-G11-SPC). The scalp was removed and scored before holes were drilled (AP +0.3, ML ±0). Cannula guides were kept in place using superglue. The skull was then covered with dental cement (GC FujiCEM 2) and dummy cannulas (Plastics One, #C315DC-SPC) were inserted into the guides. Mice were returned to their home cages and left to recover for at least 1 week.

Optical ferrule implantation. We expressed GCaMP6f selectively in dCA2 pyramidal neurons before implanting a 400-μm optic fibre, either above the dCA2 injection site in the hippocampus (4 mice, Fig. 5a, b) or over the site of dCA2 projections in dLS (5 mice, Fig. 5d, e). Mice were implanted with an optical ferrule extending 2 mm below the pedestal for LS and 1.5 mm below the pedestal for hippocampus (Doric Lenses). The scalp was removed and scored before a hole was drilled (AP +0.3, ML ±0 for LS and AP –2, ML +2.5 for hippocampus). Ferrules were kept in place using superglue. The skull was then covered with dental cement (GC FujiCEM 2). Mice were returned to their home cage and left to recover for at least 1 week.

Immunohistochemistry. Mice were transcardially perfused using saline then 4% PFA in PBS. The brains were quickly extracted and incubated in 4% PFA overnight. After 1 h washing in 0.3% glycine in PBS, 60-μm sections were prepared using a Leica VT1000S vibratome. After fixation, sections were permeabilized and blocked for 2 h with 5% goat-serum and 0.5% Triton-X in PBS at room temperature (RT). Unless indicated otherwise, sections were incubated overnight with primary antibodies at 4 °C, diluted in 5% goat-serum and 0.1% Triton-X in PBS. The sections were washed three times for 15 min in PBS and secondary antibodies were applied at RT for 3 h in in 5% goat-serum and 0.1% Triton-X in PBS. All secondary antibodies were produced in the goat, purchased from ThermoFisher Scientific and diluted at 1:500. DAPI (ThermoFisher Scientific, #D1306) staining was applied at 1:1,000 for 10 min in PBS at RT before mounting the section using fluoromount (Sigma-Aldricht). Images were acquired using an inverted confocal microscope (Leica, LSM 700).

For GFP and rabies–mCherry labelling, the first incubation was performed with chicken anti-GFP (1:1,000, AVES Labs, #GFP-1020, RRID:AB_10000240) and rabbit anti-RFP (1:500, Rockland, #600-401-379). The secondary incubation was performed with anti-chicken conjugated to Alexa 488 (#A11039, RRID:AB_142924) and anti-rabbit conjugated to Alexa 568 (#A11011, RRID:AB_143157).

For PCP4, mCherry and RGS14 labelling, the first incubation was performed with mouse IgG2a anti-RGS14 (1:50, UC Davis/NIH NeuroMab Facility, #73-170, RRID:AB_10698026) and rabbit anti-PCP4 (1:200, Sigma-Aldrich, #HPA005792, RRID:AB_1855086). The secondary incubation was performed with anti-mouse IgG2a conjugated to Alexa 488 (#A21131, RRID:AB_2535771) and anti-rabbit conjugated to Alexa 633 (#A21070, RRID:AB_2535731). We did not stain for the endogenous mCherry signal.

For Nissl, CTB-488 and PCP4 labelling, the first incubation was performed with rabbit anti-PCP4 (1:200, Sigma-Aldrich, Cat# HPA005792 RRID:AB_1855086). The secondary incubation was performed with Neurotrace 435/455 (Nissl, 1:200, #N21479, RRID:AB_2572212) and anti-rabbit conjugated to Alexa 568 (#A11011).

For Nissl, GABA and rabies–mCherry labelling, the first incubation was performed with guinea-pig anti-GABA (1:50, Abcam #ab17413). The secondary incubation was performed with Neurotrace 435/455 (Nissl, 1:200, #N21479) and anti-guinea-pig conjugated to Alexa 568 (#A11075, RRID:AB_141954).

For Nissl and c-Fos labelling, the first incubation was performed with rabbit anti-c-Fos (1:2,000, Santa Cruz, #sc52, RRID:AB_2106783) for 4 days at 4 °C. The secondary incubation was performed with Neurotrace 640/660 (Nissl, 1:200, #N21483, RRID:AB_2572212) and anti-rabbit conjugated to Alexa 488 (#A11008, RRID:AB_143165).

For mCitrine and miniRuby labelling, the first incubation was performed with chicken anti-GFP (1:1,000, AVES Labs, #GFP-1020). The secondary incubation was performed with anti-chicken conjugated to Alexa 488 (1:500, Thermo Fisher Scientific, Cat# A11039 RRID:AB_142924).

For post-hoc immunocytochemistry after patch-clamp recordings, slices were fixed for 1 h in 4% PFA in PBS. The procedure was as described above. Streptavidin conjugated to Alexa 647 (1:500, ThermoFisher Scientific, #S21374, RRID:AB_2336066) and the primary antibody anti-GFP conjugated to Alexa-488 (1:500, ThermoFisher Scientific, #A21311, RRID:AB_221477) were applied overnight at 4 °C following blocking and permeabilization.

iDISCO brain. Brains were processed as described[44]. We used the primary antibody chicken anti-GFP (1:2,000, AVES Labs, #GFP-1020) for 7 days and then the secondary antibody donkey anti-chicken conjugated to Alexa-647 (1:2,000, ThermoFisher Scientific, #A21447, RRID: AB_2535864) for 7 days. Imaging was performed using the UltraMicroscope II light-sheet microscope (LaVision). 3D reconstruction was done using Imaris software (Bitplane).

Slice preparation. For LS recordings, mice were killed under isoflurane anaesthesia by perfusion into the right ventricle of an ice-cold solution containing the following (in mM): 10 NaCl, 195 sucrose, 2.5 KCl, 10 glucose, 25 $NaHCO_3$, 1.25 NaH_2PO_4, 7 Na pyruvate, 1.25 $CaCl_2$, and 0.5 $MgCl_2$. The skull was placed in the same ice-cold medium, the brain was removed carefully from the skull and the cerebellum was cut. The brain was then glued upright with the dorsal side facing the blade and a small block of 4% agar was placed against the ventral side for mechanical stabilization. Four-hundred-micrometre coronal slices were prepared with a vibratome (VT1200S, Leica) in the same ice-cold dissection solution. Brain slices were then transferred to a chamber containing 50% dissecting solution and 50% ACSF (in mM: 125 NaCl, 2.5 KCl, 22.5 glucose, 25 $NaHCO_3$, 1.25 NaH_2PO_4, 3 Na pyruvate, 1 ascorbic acid, 2 $CaCl_2$ and 1 $MgCl_2$). The chamber was kept at 34 °C for 30 min and then at room temperature for at least 1 h before recording. All experiments were performed at 33 °C. Dissecting and recording solutions were both saturated with 95% O_2 and 5% CO_2, pH 7.4. For CA2 and CA1 pyramidal neurons recordings, transverse hippocampal slices were prepared as described[45].

Electrophysiological recordings. Slices were mounted in the recording chamber under a microscope. Recordings were acquired using a Multiclamp 700A amplifier (Molecular Device), data acquisition interface ITC-18 (Instrutech) and the Axograph X software. Whole-cell current-clamp recordings were obtained from LS cells with a patch pipette (4–5 MΩ) containing the following (in mM): 135 K methylsulfate, 5 KCl, 0.2 EGTA-Na, 10 HEPES, 2 NaCl, 5 ATP, 0.4 GTP, 10 phosphocreatine, and 5 μM biocytin, pH 7.2 (280–290 mOsm). The liquid junction potential was 1.2 mV and was not corrected. Voltage-clamp recordings were performed with an intracellular solution containing 135 Cs methylsulfate instead of K methylsulfate. Series resistance (15–25 MΩ) was monitored throughout each experiment; cells with a >20% change in series resistance were discarded. For light stimulation, pulses of blue light (pE-100, Cool LED) were delivered through a 40× immersion objective and illuminated an area of 0.2 mm^2. The illumination

© 2018 Springer Nature Limited. All rights reserved.

field was centred over the recorded cell. In a subset of experiments, the following drugs were used at the following concentrations via bath application (all drugs from Tocris unless indicated otherwise): SR 95531 (1 μM, #1262), CGP 55845 (2 μM, #1248), D-AP5 (50 μM, #0106), CNQX (20 μM, #1045), AVP ([Arg8]-vasopressin, 100 nM, #2935), dVP ([dLeu^4,Lys^8]-Avp, 50 nM, #3127), D3PVP ([deamino-Cys1, D-3-(pyridyl)-Ala^2,Arg^8-Avp], 50 nM, Sigma-Aldrich #V2257), SSR149415 (10 nM, Axon Medchem #1114), T-GOT ([Thr^4,Gly^7]-Oxt, 250 nM, Sigma-Aldrich #O6380) and CNO (5 μM, #4936). Drugs were bath applied following dilution into the external solution from stock solutions.

Data analysis for electrophysiology. A baseline recording was acquired for 10 min and then drug was applied for 15 min before measuring the effect of the drug for another 10 min. We used Axograph X software for data acquisition, and Excel (Microsoft) and PRISM (Graphpad) for data and statistical analysis. Wilcoxon or Mann–Whitney tests were performed with PRISM for statistical non-parametric comparisons of paired or non-paired data respectively. Results presented in the text and figures are reported as the mean ± s.e.m.

Social aggression. Amigo2-Cre mice and their wild-type littermates were injected with rAAV expressing iDREADD and then some were implanted with cannula guides 3 weeks later (see surgeries section above). After one week of recovery the mice were housed singly for one week before being subjected to the social aggression test.

The resident–intruder paradigm was used to assess social aggression as previously described[3,24,39]. Subject male mice (residents) were individually housed for a minimum of 1 week, with a cage change no less than 1 week before the encounter with a novel intruder. Stimulus mice (male BALB/cJ intruders) were group housed and used for only a single encounter per day. No intruder was used for more than three aggressive episodes. Experiments began at the start of the dark cycle. Feeding and water apparatuses were removed before habituation to allow unimpeded interaction and better recording. Ten-minute presentations of age- and weight-matched intruders occurred in the home cages of the resident mice after a one-hour habituation to the behavioural room. In accordance with Columbia IACUC rules, the attack was allowed to continue for 2 min after its onset, which was defined by a bite. On the rare occasions on which a stimulus mouse attacked the resident, the trial was halted and this intruder was excluded from the study. To increase the occurrences of aggression in order to enable us to quantify relevant parameters for each group, subjects were presented with up to three intruders, one each on three consecutive days. Once a subject displayed attack it was infused with miniRuby to control for the location of the drug delivery and processed for immunohistochemistry.

For intraperitoneal injection, mice were injected with CNO (10 mg/kg in saline) or vehicle (saline) 30 min before testing. For LS infusion, mice were placed under light isoflurane anaesthesia (2%) and the dummy cannula was removed. A cannula (Plastics One, #C315I-SPC) projecting 1.2 mm from the tip of the cannula guide was mounted. One microlitre of a 1 mM CNO solution (Fig. 2d), 1 μl of a 2 μM SSR149415 solution or 1 μl of saline (Fig. 6d) was infused over 5 min using a syringe pusher (Fusion 200, Chemix Inc.) mounted with a 2-μl syringe (Hamilton, #88511). The cannula was removed 2 min after the end of the micro-infusion to avoid pulling out the drug when removing the cannulas. Mice typically recovered fully from the light anaesthesia within 5 min. Mice were returned to their home cages 20 min before the test began. All encounters were recorded under red-light and sound-attenuated conditions with a Sony camera for later ethological analysis using the ANY-maze software (Stoelting Company).

Ethological analysis of aggression was performed by a blinded observer in the 2 min following the first biting attack. We measured: (1) the duration of attack within 2 min of the initial aggression; (2) the number of bites; (3) the number of tail rattles; and (4) the number of aggressive bouts. Operational definitions for these behaviours are as follows: the initiation of attack is defined by the first clear bite initiated by the resident mouse, not including mounting, excessive allogrooming, and pursuing behaviour. The duration of attack includes biting, pursuing, mounting, and excessive grooming behaviour. Attack bouts are cycles of initiated attack with continuous orientation and physical interaction by the resident towards the intruder. They are defined as completed when the resident has physically reoriented away from the intruder. The initiation of social dominance excludes biting and is defined as mounting behaviour or persistent face allo-grooming. Chi-square (χ^2) tests were performed to evaluate the statistical significance of differences in occurrences of the different behaviours. To analyse the data presented in Fig. 2b, we first performed χ^2 tests between the three control groups before pooling the control data and performing a χ^2 test between it and the test group (Cre + CNO). In a similar fashion, we used Mann–Whitney tests to analyse the data presented in Extended Data Fig. 6 by first comparing the control groups and then pooling control groups to compare them against the test group.

c-Fos experiment: behavioural paradigm. We injected saline or CNO into both Amigo2-Cre mice expressing iDREADD in dCA2 and wild-type littermates 30 min before performing the resident–intruder test. Any mouse that showed a biting attack was killed 1 h after the end of the 10-min test. We restricted our analysis to mice that showed aggression to rule out the possibility that any decrease in c-Fos expression upon silencing dCA2 could be a simple consequence of the behavioural effect of decreased aggression, rather than a result of any direct effect of silencing the dCA2–LS pathway on VMHvl activity. Because only a fraction of resident mice displayed aggression, we used the following protocol to obtain sufficient mice for analysis. Any resident that did not attack its intruder was returned it to its home cage, and then tested in the resident–intruder paradigm again 5 days later. This was repeated until the resident performed an attack, at which point it was killed for inclusion. Because CA2 silencing in Cre + CNO mice led to a decrease in the fraction of mice that displayed aggression in any one test (Fig. 2), these mice had to be subjected to more resident–intruder tests than mice in the three control groups before an attack was observed (Extended Data Fig. 10c). However, we found that repeated tests separated by a 5-day interval did not alter the incidence of aggression in control mice (33% in first test versus 35% in second test), suggesting that there were no cumulative behavioural effects of the repeated testing (as long as a 5-day inter-test interval was used).

Data analysis for c-Fos$^+$ cell counting in the VMHvl. For each mouse, we randomly selected two non-consecutive 60-μm thick coronal sections between 1.4 and 1.9 mm from Bregma along the rostral–caudal axis and stained them for c-Fos and Nissl. High-resolution 16-μm stacks of the hypothalamus were acquired and projected along the z-axis using a LSM 700 confocal microscope (Zeiss). We identified the VMH based on Nissl staining and hypothalamic hallmarks (fornix, third ventricle). Additionally, we performed VGlut2 immunostaining on a limited number of slices from wild-type mice to confirm the location of the VMH[3]. We identified the VMHvl subnucleus as consisting of the ventral third of the VMH, and manually counted c-Fos$^+$ cells in this region, making sure they co-localized with Nissl staining. We verified that the total surface area analysed was similar between mice. Results were averaged across bilateral regions and sections for each mouse.

Novel environment and novel object exploration. Isolated Amigo2-Cre mice and their wild-type littermates previously injected with rAAV expressing iDREADD were given 10 mg/kg CNO intraperitoneally 30 min before being introduced into a new arena (60 cm × 60 cm). They were allowed to roam freely for 10 min. Subsequently a novel object (pen) was introduced in the middle and they were allowed to explore the object for another 10 min. The session was recorded using a video camera (Imaging Source) and tracked online using the AnyMaze 7 software (Stoelting). Offline analysis measured the total distance travelled during the first 10 min as well as centre–surround preference. We also used AnyMaze 7 to measure the time spent investigating the novel object. Mann–Whitney tests were performed to compare the two groups.

Novel mouse exploration (sociability). Isolated Amigo2-Cre mice and their wild-type littermates previously injected with rAAV expressing iDREADD were given 10 mg/kg CNO intraperitoneally 30 min before being presented to an intruder (see above). All resident mice used in the analysis displayed an initial period of social exploration of the intruder during the resident–intruder test, and were scored offline for the time spent interacting with the intruder mouse.

Prey aggression. Isolated Amigo2-Cre mice and their wild-type littermates previously injected with rAAV expressing iDREADD were given 10 mg/kg CNO intraperitoneally 30 min before a live cricket was introduced into their home cage. We measured the latency to attack and whether they did or did not attack the prey. Mice were food-deprived for 12 h before the experiment.

Feeding. Isolated mice implanted for fibre photometry were food-deprived for 12 h. A food pellet was introduced into their home cage and we recorded CA2 activity during 10 min of feeding.

Female interaction. Isolated mice implanted for fibre photometry were presented with a female in oestrus in their home cage for 15 min. Oestrus was induced in ovariectomized females (C56Bl/6J) as described[46]. In brief, gonadectomized, steroid-primed C57Bl/6J females (implanted with a capsule containing 50 μg oestradiol benzoate in 25 μl sesame oil, followed by a subcutaneous injection of 0.5 mg progesterone in 25 μl sesame oil 4–6 h before use) were used as stimulus mice. An implant of progesterone was inserted into their neck and an intraperitoneal injection of oestradiol was given 4 h before the test.

Multiple ovariectomized female interaction. Isolated mice implanted for fibre photometry were presented with a gonadectomized C57Bl/6J female mouse for 5 min in the test mouse home cage. The presentation was repeated four times at 10-min intervals. Upon the fifth presentation, a novel ovariectomized female was presented for 5 min[13].

Fibre photometry recordings. Fibre photometry was conducted in a similar way to previous studies[31,47,48]. Two LEDs (405 nm and 473 nm) were coupled to a fluorescence mini-cube (FMC) and 1 × 1 fibre optic rotary joint to deliver light into optical fibres permanently implanted above the lateral septum or CA2 during behaviour. Emitted light between 420 and 450 nm (with 405 nm excitation) and 500 and 540 nm (with 473 nm excitation) were collected through the FMC on separate fibre-coupled Newport 2151 photo-receiver modules. The collected fluorescent signals were collected in AC-high mode and converted to voltage via the

© 2018 Springer Nature Limited. All rights reserved.

formula $V = PRG$, where V is collected voltage, P is the optical input power in watts, R is photodetector responsivity in amps/watts (0.2 – 0.4), and G is the trans-impedance gain of the amplifier. Raw signals for 473 nm excitation (GCaMP6f) and 405 nm excitation (background auto-fluorescence) were recorded and processed using Doric Neuroscience Studio software. Subtraction of the background fluorescence was calculated via a time-fitted running average of the 473 nm channel relative to the 405 nm control channel and normalized by the 405 nm signal using the formula (473 nm – 405 nm)/405 nm. Finally, a peak enveloping Fourier transform was applied to the $\Delta F/F$ signal across the entire trace to identify peaks in activity. Light was delivered at a final intensity of 2.24 mW (473 nm) and 2.76 mW (405 nm) at the tip of the patch-cord before coupling with the implanted ferrule. Mice were habituated to the fibre for three days by placing the fibre on the mouse head and letting it roam free for 1 h. Mice were also housed with a female for one night in order to increase their aggression. On the fourth day, we conducted the resident–intruder test as described above and recorded the interaction for 10–15 min. Other behavioural tests were conducted on the following days. We measured peak and mean fluorescence during each behavioural episode and normalized them by the average of the peak or mean fluorescence in between each interaction episode while the mouse was freely moving in its cage (non-social exploration of the cage). Mouse tracking using AnyMaze was used to calculate mouse velocity. Pearson correlation coefficient was used to calculate the correlation between fluorescent signal and velocity.

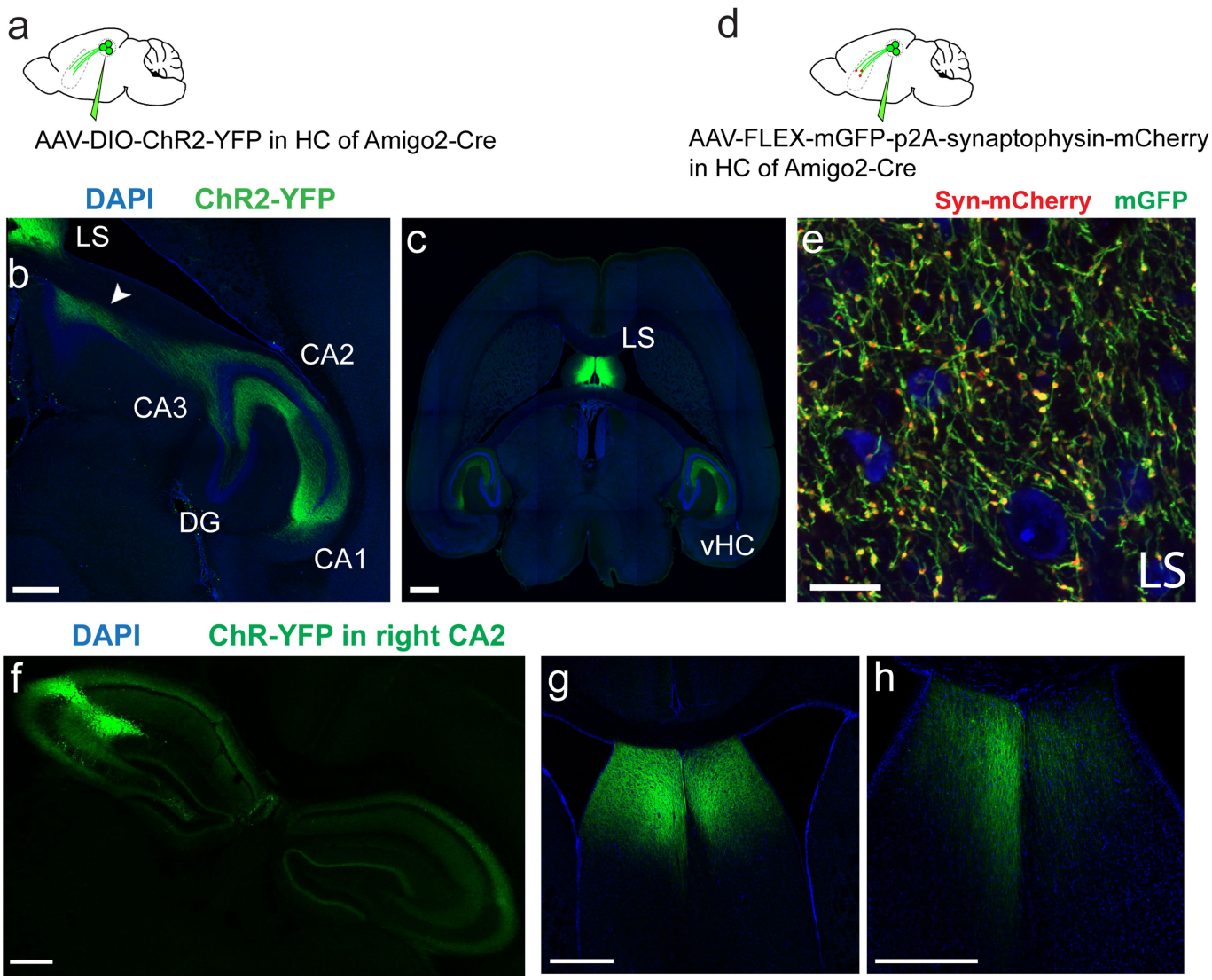
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All analysed data supporting this study are presented in the form of graphs. All raw records used in the analysis are available from the corresponding author in response to reasonable requests.

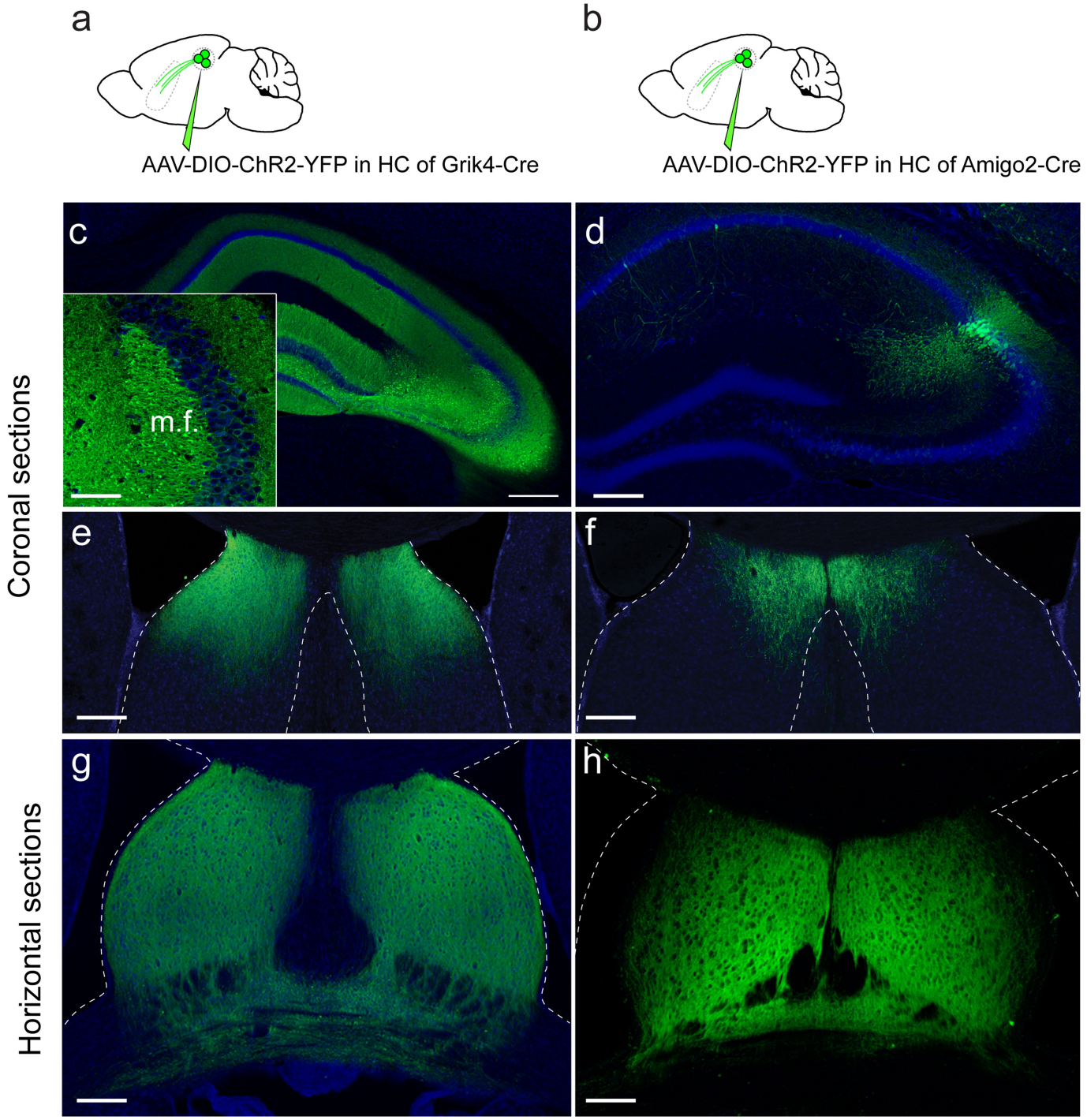
38. Nakazawa, K. et al. Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science* **297**, 211–218 (2002).
39. Wersinger, S. R., Ginns, E. I., O'Carroll, A.-M., Lolait, S. J. & Young, W. S., III. Vasopressin V1b receptor knockout reduces aggressive behavior in male mice. *Mol. Psychiatry* **7**, 975–984 (2002).
40. Berndt, A. et al. High-efficiency channelrhodopsins for fast neuronal stimulation at low light levels. *Proc. Natl Acad. Sci. USA* **108**, 7595–7600 (2011).
41. Urban, D. J. & Roth, B. L. DREADDs (designer receptors exclusively activated by designer drugs): chemogenetic tools with therapeutic utility. *Annu. Rev. Pharmacol. Toxicol.* **55**, 399–417 (2015).
42. Chen, T.-W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
43. Callaway, E. M. & Luo, L. Monosynaptic circuit tracing with glycoprotein-deleted rabies viruses. *J. Neurosci.* **35**, 8979–8985 (2015).
44. Renier, N. et al. Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* **165**, 1789–1802 (2016).
45. Leroy, F., Brann, D. H., Meira, T. & Siegelbaum, S. A. Input-timing-dependent plasticity in the hippocampal CA2 region and its potential role in social memory. *Neuron* **95**, 1089–1102.e5 (2017).
46. Kikusui, T. in *Pheromone Signaling. Methods in Molecular Biology* (ed. Touhara, K.) 307–318 (Humana, Totowa, 2013).
47. Lerner, T. N. et al. Intact-brain analyses reveal distinct information carried by SNc dopamine subcircuits. *Cell* **162**, 635–647 (2015).
48. Gunaydin, L. A. et al. Natural neural projection dynamics underlying social behavior. *Cell* **157**, 1535–1551 (2014).
49. Kohl, J. et al. Functional circuit architecture underlying parental behaviour. *Nature* **556**, 326–331 (2018).
50. Watson, C. & Paxinos, G. *Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates* 4th edn (Academic, 2012).
51. Swanson, L. W., Sawchenko, P. E. & Cowan, W. M. Evidence that the commissural, associational and septal projections of the regio inferior of the hippocampus arise from the same neurons. *Brain Res.* **197**, 207–212 (1980).

© 2018 Springer Nature Limited. All rights reserved.



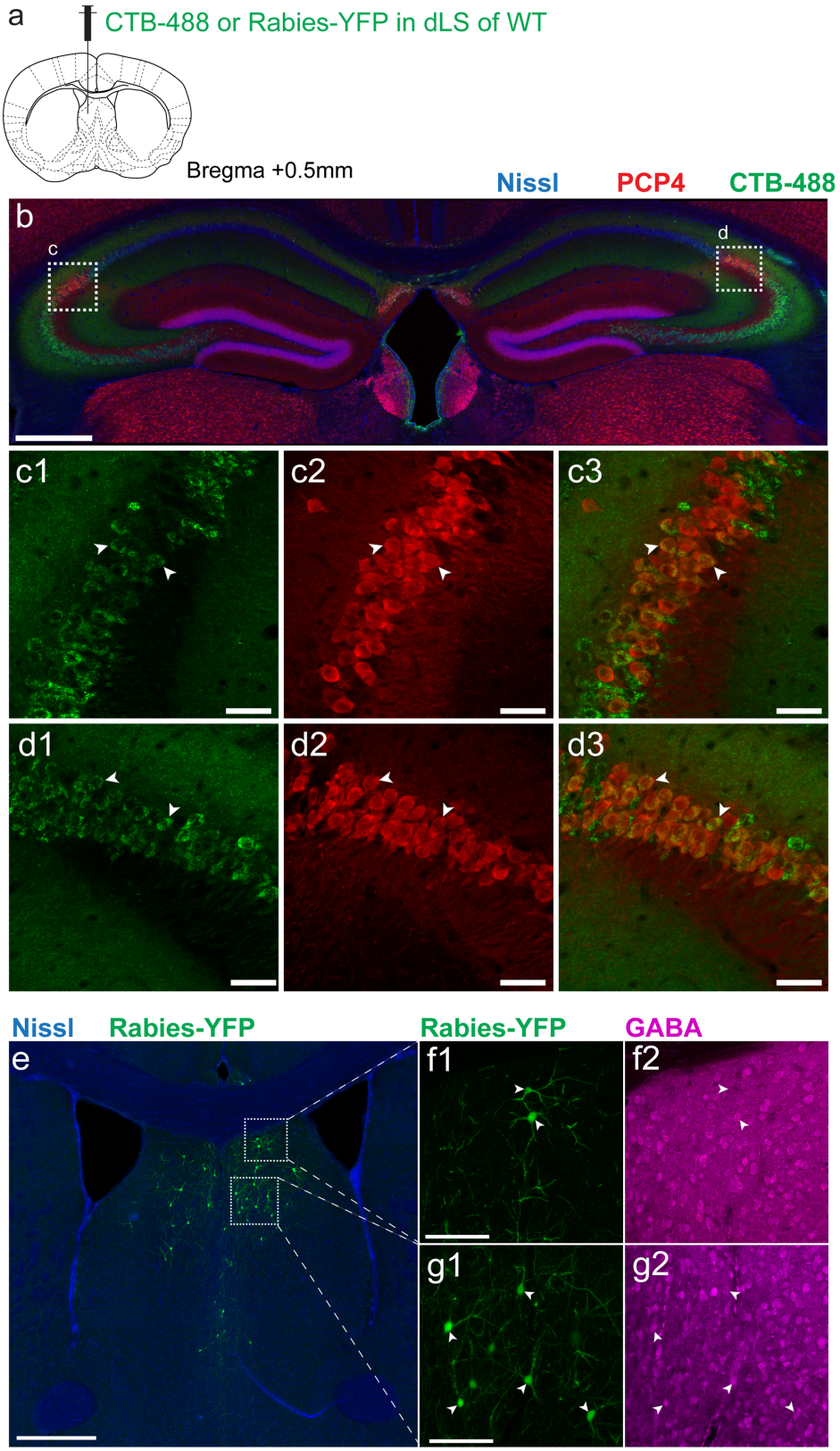
Extended Data Fig. 1 | CA2 projections to LS. a–c, Diagram and horizontal sections from an Amigo2-Cre mouse brain injected with rAAV5-EF1a-DIO-hChR2(E123T/T159C)-eYFP into dCA2. The arrowhead in **b** shows the dCA2 axons extending up to the dLS. Drawing in **a** was inspired by ref. [49]. **c**, More ventral, showing the projection from dCA2 to LS and to the ventral hippocampus. **d**, **e**, Diagram and enlarged view of a coronal dLS section from an Amigo2-Cre mouse brain injected into dCA2 with rAAV5-mGFP-p2A-synatophysin-mCherry, labelling dCA2 projections (green) and presynaptic terminals (red). **f–h**, Coronal sections of an Amigo2-Cre mouse brain injected unilaterally with rAAV5-EF1a-DIO-hChR2(E123T/T159C)-eYFP into the right CA2 area. **f**, Hippocampal section. **g**, **h**, LS sections. Three mice were injected per experiment. All mice presented similar staining patterns. Scale bars, 500 μm except **e**, 20 μm.

© 2018 Springer Nature Limited. All rights reserved.



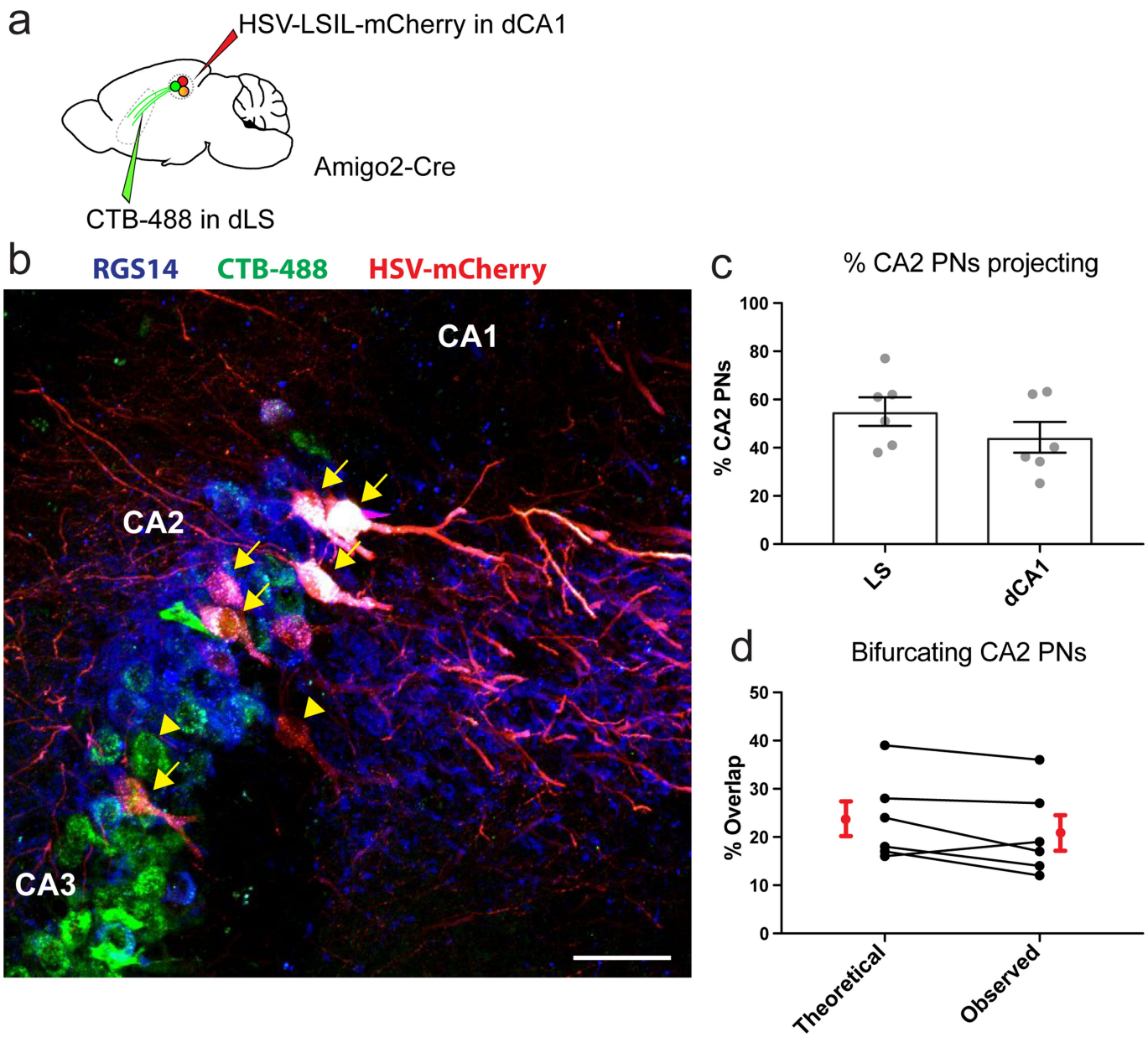
Extended Data Fig. 2 | dCA3 and dCA2 project to different but overlapping regions of LS. a, **b**, rAAV5-EF1a-DIO-hChR2(E123T/T159C)-eYFP was injected into the hippocampal dCA3 region of a Grik4-Cre mouse (**a**) and the dCA2 region of an Amigo2-Cre mouse (**b**). **c**–**f**, Hippocampal (**c**, **d**) and LS coronal (**e**, **f**) sections. **g**, **h**, Horizontal LS sections. m.f., mossy fibre. Three mice were injected per experiment. All mice presented similar staining patterns. Scale bars, 200 μm except inset in **c**, 50 μm.

© 2018 Springer Nature Limited. All rights reserved.



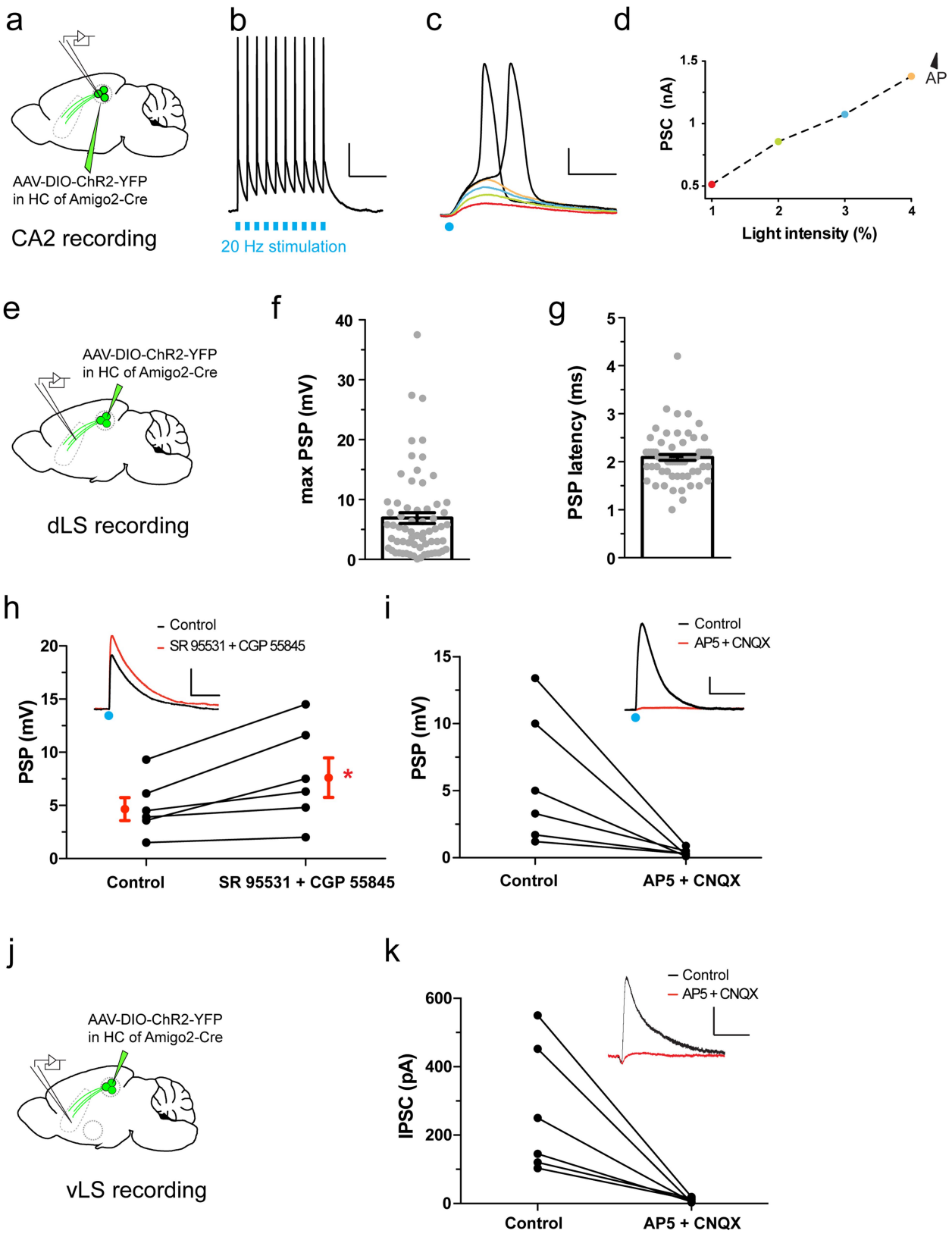
Extended Data Fig. 3 | CTB and rabies virus injections in dLS confirm the CA2–LS projection. a, Atlas drawing (reproduced with permission[50]) of the CTB-488 injection site. **b**–**d**, Hippocampal coronal sections following injection of CTB (green) into right LS and labelled for PCP4 (red); blue, DAPI staining. **b**, Whole hippocampi. **c**, **d**, High-magnification views of CA2 region in ipsilateral (**c**) and contralateral (**d**) hippocampus. Three mice injected. **e**, Coronal section of LS injected with G-deleted rabies virus (green). **f**, **g**, Locally infected cells (arrowheads, green) enlarged and labelled for GABA immunofluorescence (arrows, purple). Three mice were injected per experiment. All mice presented similar staining patterns. Scale bars, **b**, **e**, 500 μm, **c**, **d**, 50 μm, **f**, **g**, 100 μm.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | dCA2 pyramidal neurons project to both dCA1 and LS. a, Dual retrograde staining of dCA2 by injection of HSV-LSL1–mCherry into dCA1 and CTB-488 into dLS of an Amigo2-Cre mouse. **b**, Hippocampal coronal section following injections as in **a** and labelled for RGS14 (blue). Arrowheads denote single-labelled cells (green or red) and arrows dual-labelled ones (white). Scale bar, 50 μm. **c**, Quantification of the percentage of CA2 pyramidal neurons that project to either dLS or dCA1. Because retrograde labelling efficiency is not complete, the fraction of labelled cells provides a lower limit on the fraction of dCA2 pyramidal neurons that project to these regions (12 sections from 6 mice in each group). Bars show mean ± s.e.m. **d**, Comparison of the expected percentage of dual-labelled CA2 pyramidal neurons versus the observed percentage. The fraction of dual-labelled cells (21 ± 4%) was almost identical to the fraction predicted under the assumption of random retrograde labelling of a single population of CA2 pyramidal neurons, each of which sends a projection to CA1 and LS ([fraction of labelled CA1 projecting cells] × [fraction of labelled LS projecting cells] = 0.55 × 0.44 = 24%). This is similar to results suggesting that a uniform population of CA3 pyramidal neurons projects to both LS and CA1[51]. Thus, it is likely that a single population of CA2 pyramidal cells projects to both LS and CA1. Two-sided Wilcoxon test, $P = 0.2$. Black dots, individual mice ($n = 6$). Red dots with error bars, mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.

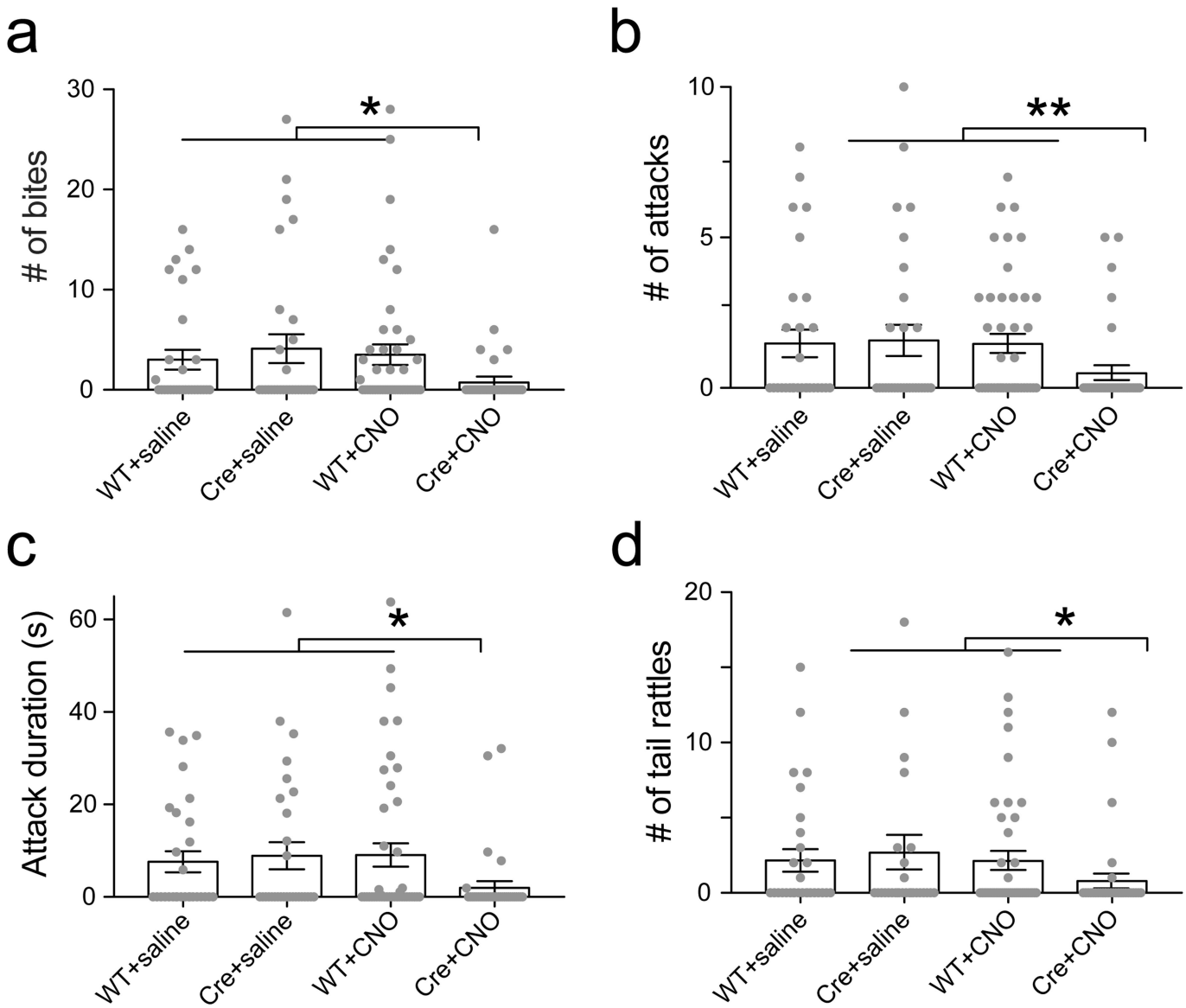


Extended Data Fig. 5 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

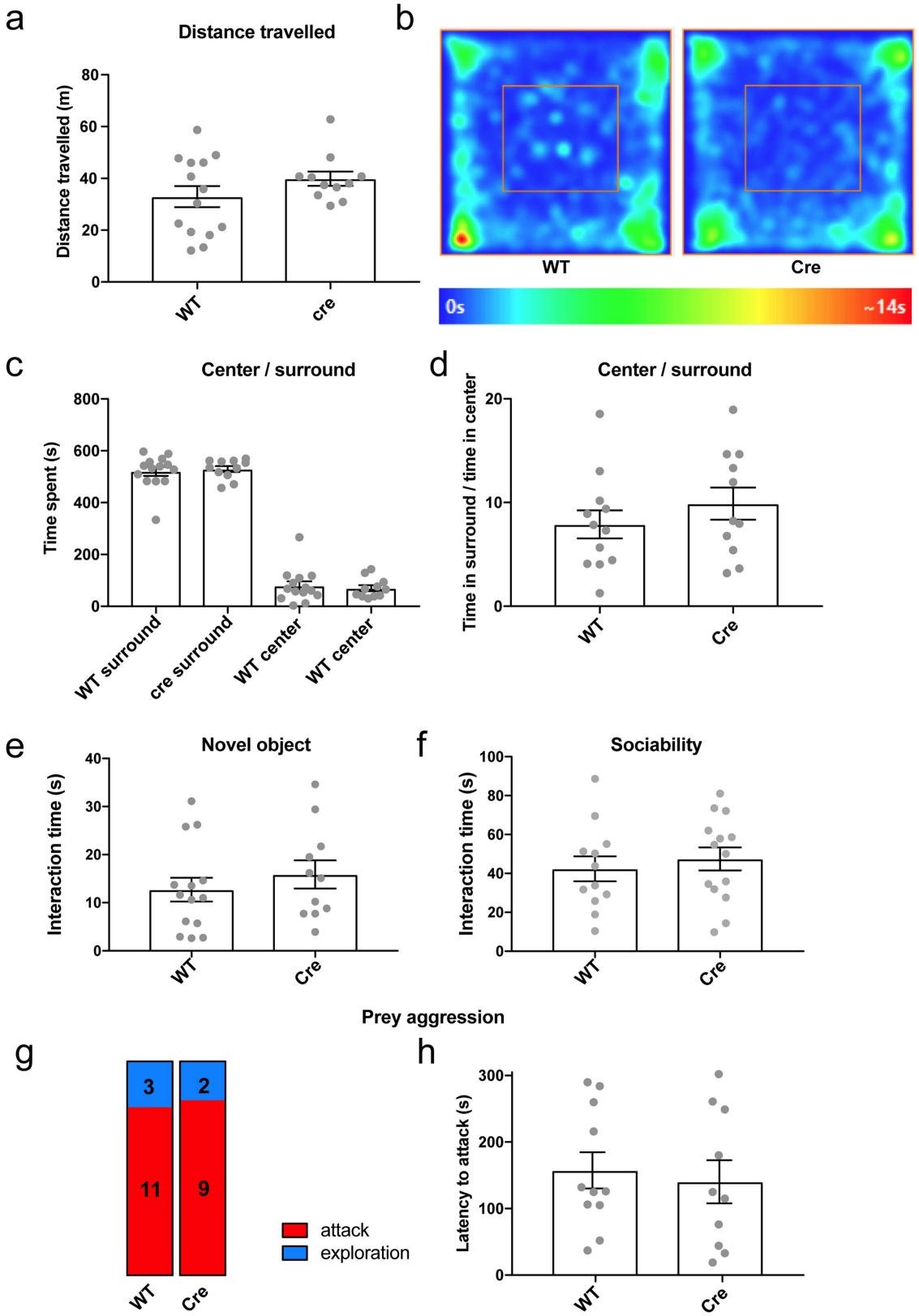
Extended Data Fig. 5 | Analysis of light-evoked synaptic responses with ChR2–eYFP expressed in dCA2 pyramidal neurons. a, Schematic of hippocampal or septal slice recordings from Amigo2-Cre mouse injected into dCA2 with rAAV5-EF1a-DIO-hChR2(E123T/T159C)-eYFP (5 cells from 3 mice). **b**, CA2 pyramidal neuron voltage response under current-clamp to a 20-Hz train of 1-ms light pulses. Scale bars, 20 mV and 200 ms. **c**, dCA2 pyramidal neuron current responses under voltage-clamp with increasing light stimulation intensity (onset at blue spot). Negative (excitatory) currents plotted upwards. Responses to different light intensities coloured as in **d**. Note action currents reflecting escaped action potentials with two highest light intensities. Scale bars, 1 nA and 5 ms. **d**, Input–output curve for light-induced current in ChR2-expressing dCA2 pyramidal neurons. **e**, Schematic of dLS synaptic voltage recordings in septal slice from Amigo2-Cre mouse injected with the same virus as in **a**. **f**, **g**, Maximal dLS PSP amplitude (**f**) and latency (**g**) following light stimulation (65 cells from 32 mice; mean ± s.e.m.). **h**, Quantification of effect of $GABA_A$ and $GABA_B$ receptor antagonist application (1 μM SR 95531 and 2 μM CGP 55845, respectively) on light-induced PSP amplitude in dLS showing individual cells (black, 6 cells from 4 mice) and mean ± s.e.m. (red). Two-sided Wilcoxon test, $^*P = 0.03$. Scale bars, 2 mV and 100 ms. **i**, Light-induced PSP amplitude in dLS before and after application of NMDA and AMPA receptor antagonists 25 μM AP5 and 20 μM CNQX, respectively (6 cells from 3 mice). Two-sided Wilcoxon test compared to baseline, $P = 0.03$. Scale bars, 5 mV and 100 ms. **j**, vLS recording under same conditions as in **a** and **e**. **k**, Amplitude of light-induced IPSCs in vLS before and after application of 25 μM AP5 and 20 μM CNQX (6 cells from 3 mice). Two-sided Wilcoxon test, $P = 0.03$. Scale bars, 200 pA and 100 ms.

© 2018 Springer Nature Limited. All rights reserved.



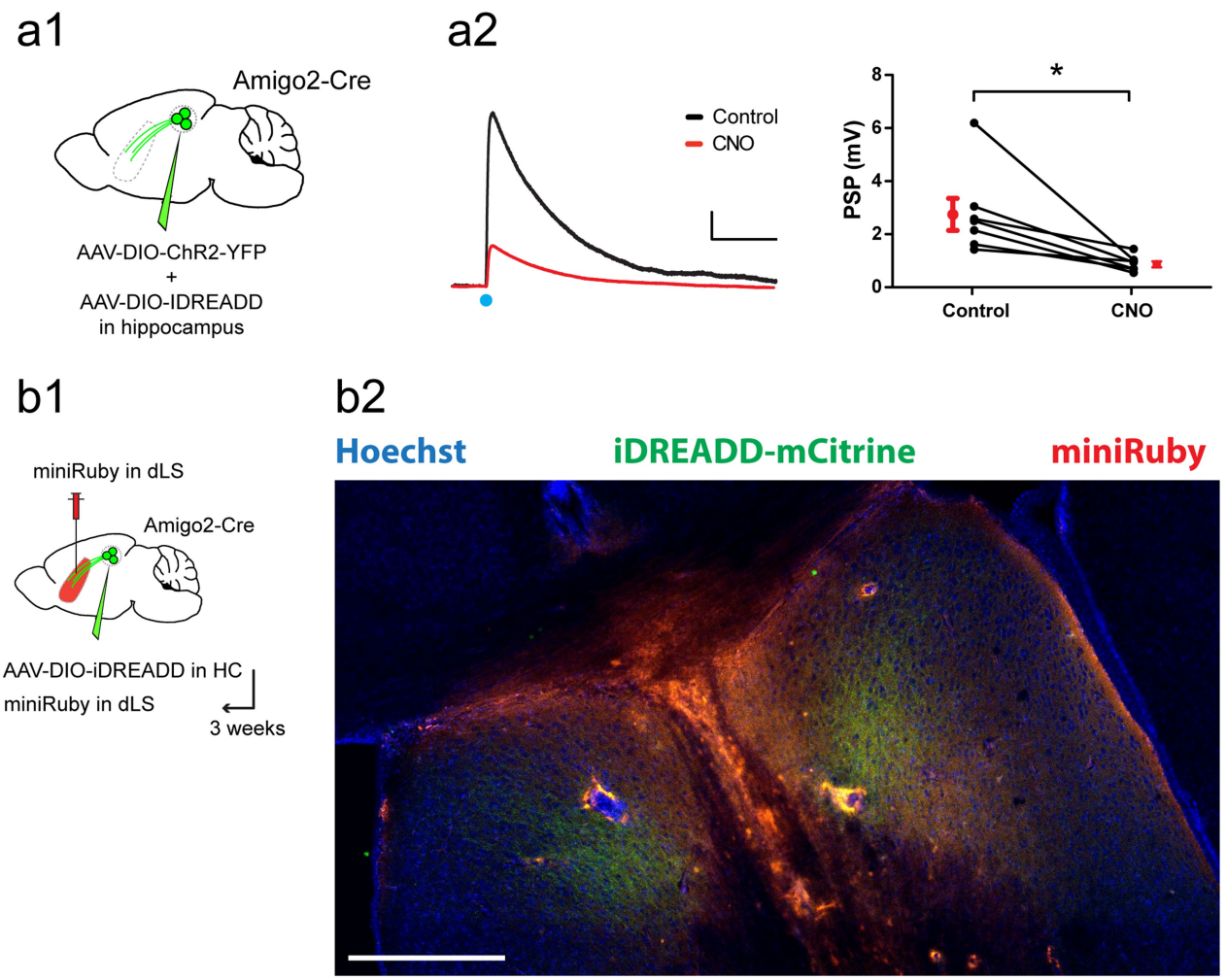
Extended Data Fig. 6 | Effect of CA2 silencing on aggression attack parameters. Amigo2-Cre (Cre) mice and wild-type littermates (WT) were injected into dCA2 with rAAV2-hsyn-DIO-HA-hM4D(Gi)-IRES-mCitrine (AAV-DIO-iDREADD). After 3 weeks, mice were injected intraperitoneally with CNO or saline and, 30 min later, subjected to the resident–intruder test. **a**–**d**, Mean (± s.e.m.) number of bites (**a**), number of attack bouts (**b**), attack duration (**c**) and number of tail rattles (**d**). Grey dots, individual mice (29, 29, 43 and 34 mice from left to right for all bar graphs). Two-sided Mann–Whitney tests: $*P = 0.011$; $**P = 0.096$; $*P = 0.012$; and $*P = 0.041$, for **a**–**d**, respectively.

© 2018 Springer Nature Limited. All rights reserved.



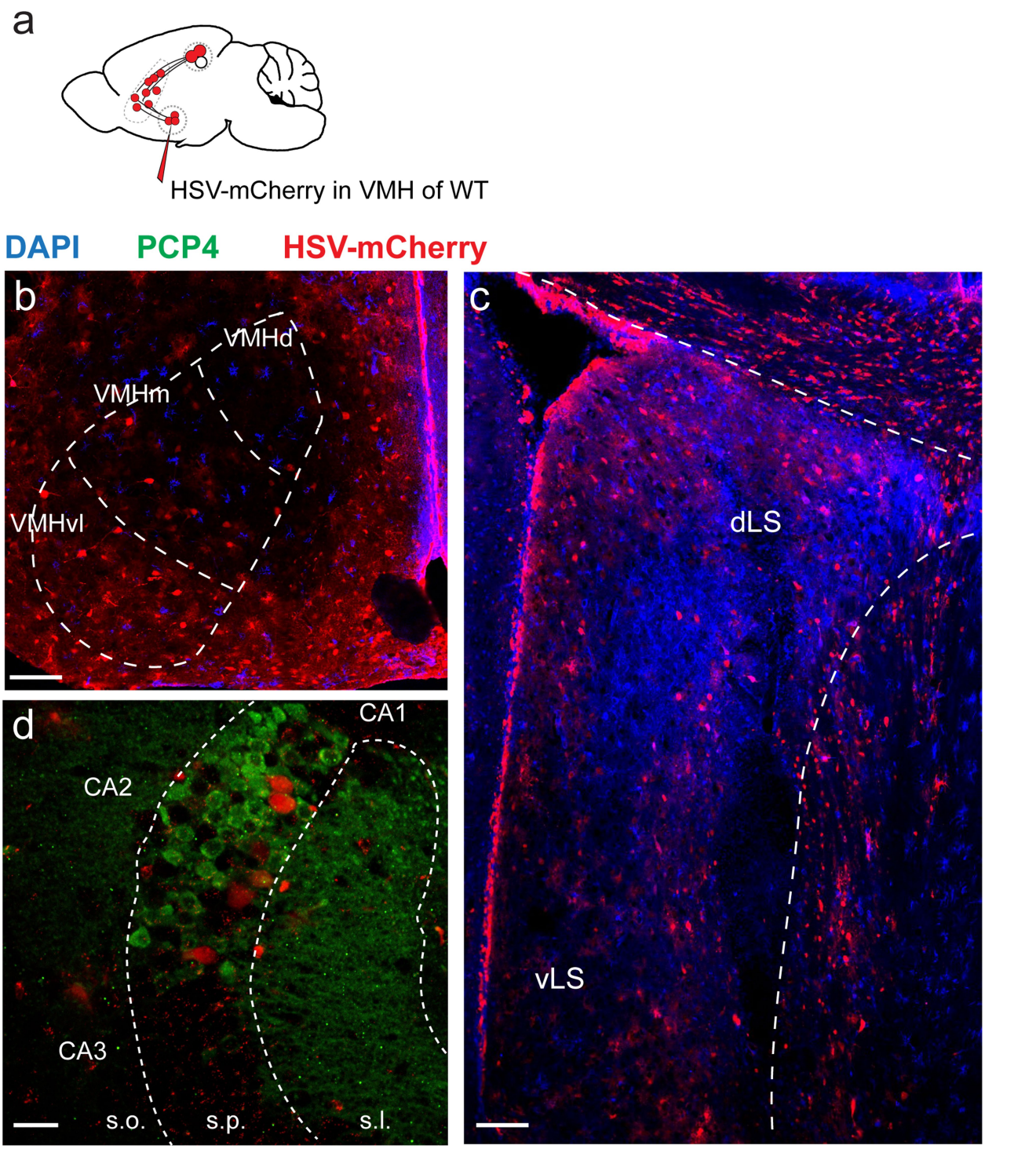
Extended Data Fig. 7 | Behavioural controls for CA2 silencing. Amigo2-Cre mice (Cre) and wild-type littermates (WT) were injected into dCA2 with rAAV2-hsyn-DIO-HA-hM4D(Gi)-IRES-mCitrine. After 3–4 weeks for viral expression, mice were given CNO (10 mg kg^{-1}) intraperitoneally 30 min before behavioural testing. **a**–**d**, Open field testing. **a**, Distance travelled. **b**, Heat maps of time spent at each position for a wild-type and an Amigo2-Cre mouse (all 25 mice showed similar heat maps). **c**, Time spent in centre and surround. **d**, Ratio of the time spent in surround/centre. **e**, Time spent interacting with a novel object. **f**, Time spent interacting with a novel mouse. **g**, Stacked bar charts of the distributions of mice that attacked or only explored the cricket in prey aggression test. **h**, Latency to attack the cricket. Grey dots, individual mice (14 wild-type and 11 Amigo2-Cre mice). Bars show mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



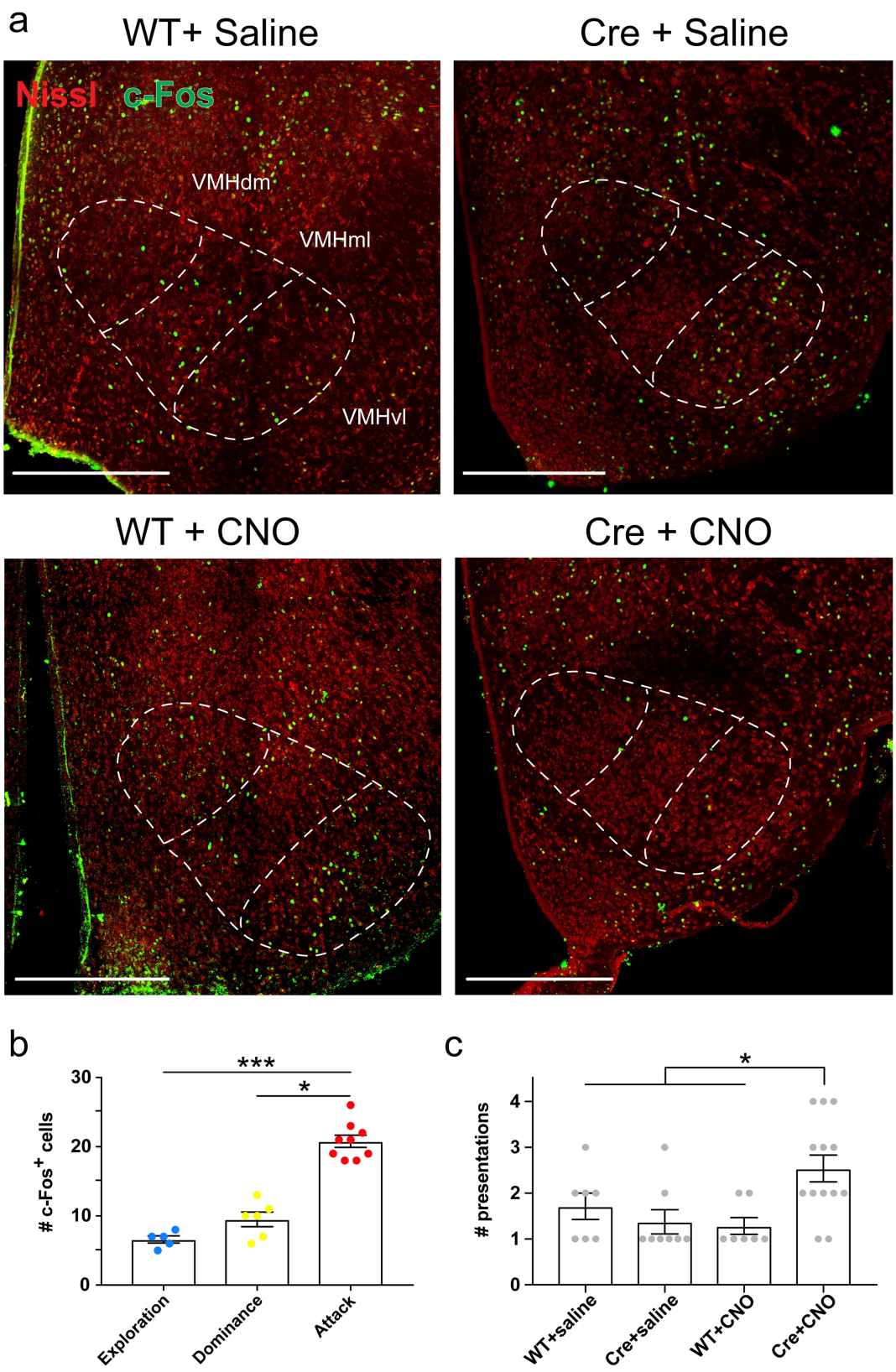
Extended Data Fig. 8 | Silencing dCA2 inputs to dLS by CNO application reduces PSP in dLS neurons evoked by photostimulation. **a1**, Co-expression of iDREADD and ChR2 in CA2 pyramidal neurons. **a2**, Left, light-evoked PSP in dLS neuron before and after bath application of 5 μM CNO. Scale bars, 1 mV and 100 ms. Right, quantification of effect of CNO on peak PSP amplitude showing individual cells (black, 7 cells from 3 mice) and mean ± s.e.m. (red). Two-sided Wilcoxon test, $*P = 0.02$. **b**, Schematic (**b1**) and LS coronal section (**b2**) from an Amigo2-Cre mouse expressing iDREADD and mCitrine in CA2 pyramidal neurons and implanted with a cannula in LS. Mouse was infused with 1 μl miniRuby through the cannula 15 min before death. Labelling shows mCitrine expression (green) and miniRuby (red). Scale bar, 400 μm.

© 2018 Springer Nature Limited. All rights reserved.



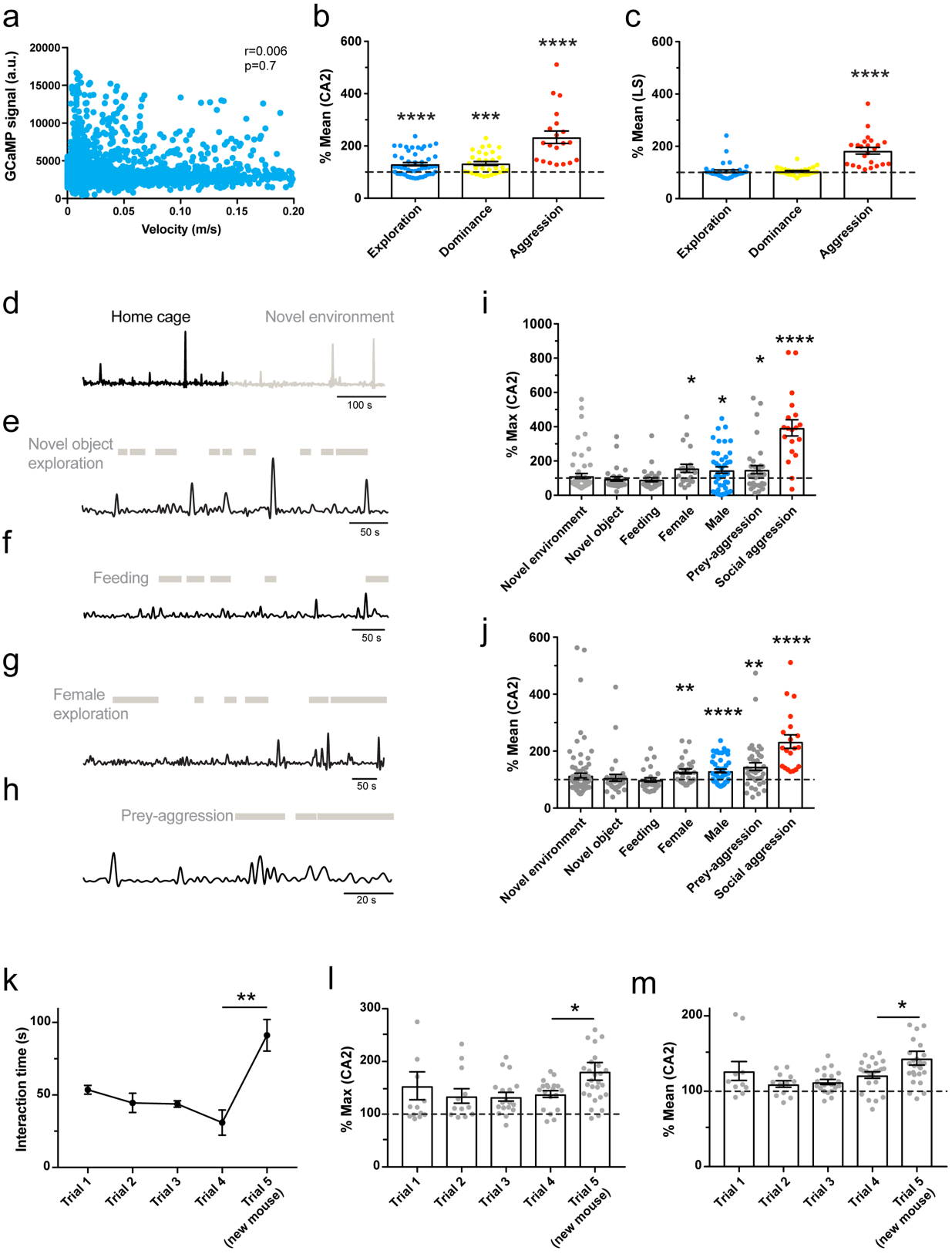
Extended Data Fig. 9 | Hippocampal retrograde labelling following HSV injection into VMH. a, Schematic of the experiment. **b**–**d**, Coronal sections after injection of the retrograde trans-synaptic tracer HSV CMV–mCherry into VMH. Note the labelling in dorsal and ventral LS (**c**) and dCA2 (**d**). Three mice injected. All mice presented a similar staining pattern. Scale bars, **b**, **c**, 100 μm, **d**, 40 μm.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | VMHvl c-Fos quantification controls. **a**, Representative images of c-Fos immunofluorescence in VMH following bouts of aggression 30 min after intraperitoneal injection of CNO or saline in wild-type or Amigo2-Cre mice (all groups previously injected in CA2 with rAAV2-hsyn-DIO-HA-hM4D(Gi)-IRES-mCitrine). Scale bars, 400 μm. **b**, Number of c-Fos-expressing cells in VMHvl of wild-type mice killed 1 h after indicated behaviours. Coloured dots, individual mice (5, 6 and 9 mice). Bars show mean ± s.e.m. Kruskal–Wallis test $\chi^2 = 15.16$, $P < 0.0001$ followed by Dunn's multiple comparison tests, $^{***}P = 0.0006$ and $^{*}P = 0.01$. **c**, Each mouse was challenged with a resident–intruder test once every 5 days until it showed aggression, after which it was killed. Grey dots, individual mice (7, 8, 7 and 13). Bars show mean ± s.e.m. Two-way ANOVA: $F = 7.4$, $^{*}P = 0.01$.

© 2018 Springer Nature Limited. All rights reserved.

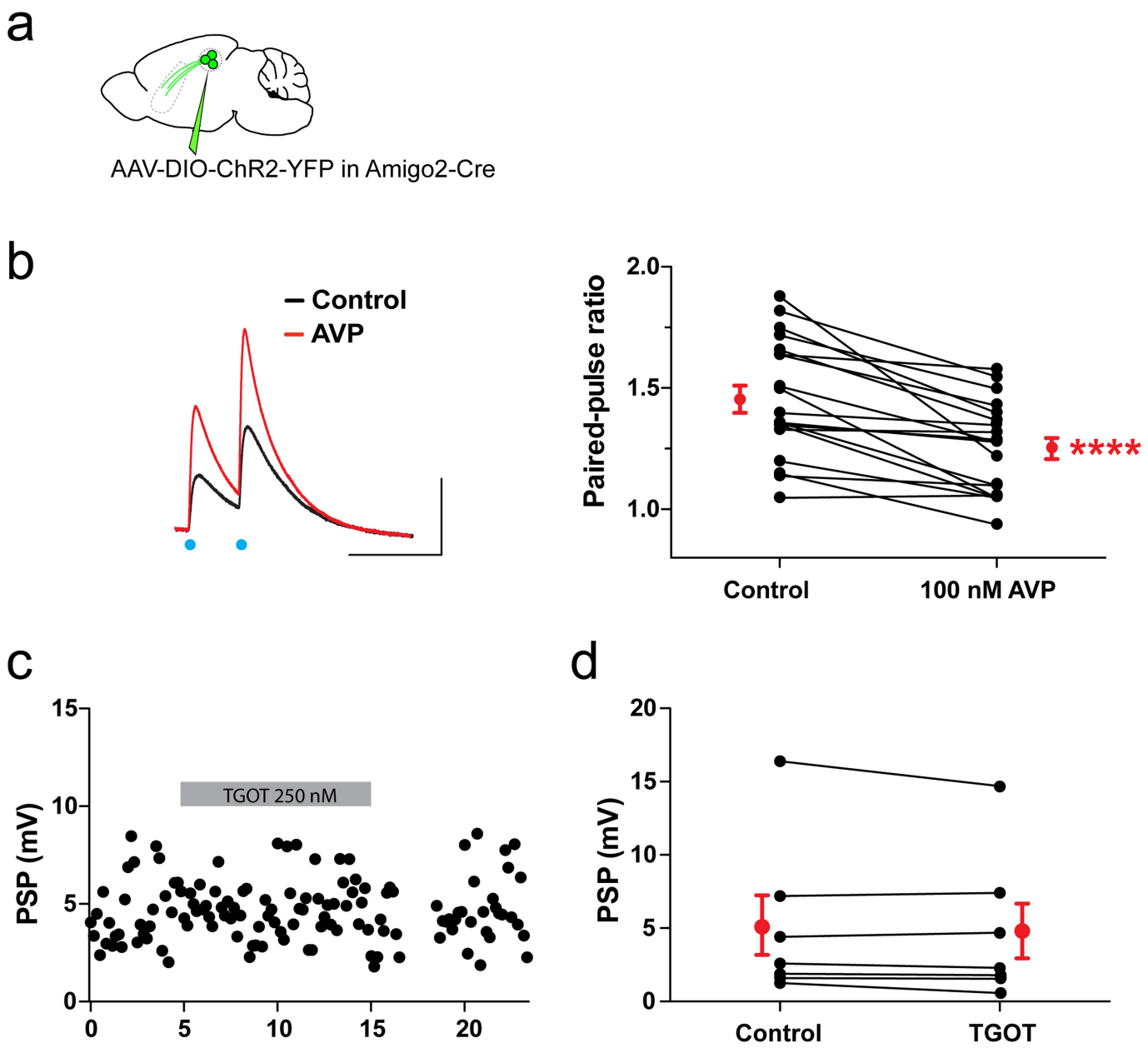


Extended Data Fig. 11 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 11 | Fibre photometry measures of CA2 pyramidal neuron GCaMP6f signals during social interactions. a, GCaMP6f fluorescence signal as a function of mouse velocity. Pearson test of correlation (r) performed on $n = 5{,}332$ 1-s time bins. Note the lack of correlation ($P = 0.7$). Lack of correlation was observed in all mice. **b**, GCaMP6f responses with fibre over CA2 pyramidal neuron soma during indicated social behaviour (calculated from the mean signal during each episode). Each point represents an episode (49, 31 and 20 episodes from left to right collected from 5 mice). Two-sided one-sample t-tests; $^{****}P < 0.001$, $^{***}P = 0.0001$. **c**, As in **b** with fibre over CA2 pyramidal neuron terminals in LS (34, 27 and 23 episodes collected from 6 mice). Bars show mean ± s.e.m. Two-sided one-sample t-test; $^{****}P < 0.001$. **d**–**j**, GCaMP6f responses from CA2 somata during various social and non-social behaviours. GCaMP6f recording from CA2 pyramidal neurons in hippocampus during: exploration of a novel environment (**d**), exploration of a novel object (**e**), feeding (**f**), exploration of a **f**emale (**g**), and predator–prey a**gg**ression (**h**). Bar graphs of the GCaMP6f responses during each type of social interaction (calculated from the maximum (**i**) or the mean (**j**) signal during each episode). Each point represents an episode except for the novel environment, where episodes were defined as 10-s bins (60, 27, 17, 27, 34, 21, 43, 19 episodes from left to right collected from 6 mice). Bars show mean ± s.e.m. Two-sided one-sample t-tests for **i**; $^{*}P = 0.03$, $^{*}P = 0.02$, $^{*}P = 0.047$ and $^{****}P < 0.0001$ from left to right. Two-sided one-sample t-tests for **j**; $^{**}P = 0.003$, $^{****}P < 0.0001$, $^{**}P = 0.002$ and $^{****}P < 0.0001$ from left to right. **k**–**m**, Behaviour and GCaMP6f responses from dCA2 pyramidal neuron somata during multiple presentations of an ovariectomized female. **k**, Time (mean ± s.e.m.) spent in social exploration of the same novel female presented in trials 1–4 and a second novel female in trial 5 (4 mice). Resident males normally engaged in non-aggressive social exploration of the female during each exposure, with exploration time showing a trend to decrease during successive trials as a result of increased familiarity. Introduction of a novel female in trial 5 resulted in enhanced social exploration, indicating that the decrease in exploration of the same female represents social memory formation, and not fatigue or disinterest. We previously found that this behaviour required dCA2[13]. Two-sided t-test, $^{**}P = 0.005$. **l**, **m**, Peak (**l**) and mean (**m**) GCaMP6f responses recorded from CA2 pyramidal neurons in hippocampus during each period of social exploration in each trial (10, 12, 20, 23 and 26 episodes collected from 4 mice; mean ± s.e.m.). Fibre photometry recordings show that dCA2 responds during social exploration of the female in the familiarization trials, with a trend for activity to decrease with increased familiarization. Introduction of the novel female produced a statistically significant increase in the dCA2 GCaMP6f signal (trial 5). Two-sided t-tests: $^{*}P = 0.048$ (**l**) and $^{*}P = 0.039$ (**m**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 12 | Neuromodulation of the CA2-LS synapse by AVP. a, ChR2 expression in dCA2 pyramidal neurons. **b**, Paired-pulse ratios (right) of PSP amplitudes evoked by two light pulses separated by 50 ms (left), before and after 100 nM AVP (19 cells from 10 mice). Black dots, individual cells; red dots, mean ± s.e.m. Two-sided Wilcoxon test, ****$P < 0.0001$. Scale bars, 2 mV, 100 ms. **c**, Example time course of PSPs in dLS evoked by photostimulation of ChR2-expressing dCA2 projections. Grey bar shows period of bath application of 250 nM TGOT. **d**, PSP amplitudes before and 30 min after application of 250 nM TGOT (7 cells from 3 mice). Black dots, individual cells; red dots, mean ± s.e.m. Two-sided Wilcoxon test, $P = 0.3$.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0744-4

Single-cell mapping of lineage and identity in direct reprogramming

Brent A. Biddy[1,2,3], Wenjun Kong[1,2,3], Kenji Kamimoto[1,2,3], Chuner Guo[1,2,3], Sarah E. Waye[1,2,3], Tao Sun[1,2,3,4] & Samantha A. Morris[1,2,3]*

Direct lineage reprogramming involves the conversion of cellular identity. Single-cell technologies are useful for deconstructing the considerable heterogeneity that emerges during lineage conversion. However, lineage relationships are typically lost during cell processing, complicating trajectory reconstruction. Here we present 'CellTagging', a combinatorial cell-indexing methodology that enables parallel capture of clonal history and cell identity, in which sequential rounds of cell labelling enable the construction of multi-level lineage trees. CellTagging and longitudinal tracking of fibroblast to induced endoderm progenitor reprogramming reveals two distinct trajectories: one leading to successfully reprogrammed cells, and one leading to a 'dead-end' state, paths determined in the earliest stages of lineage conversion. We find that expression of a putative methyltransferase, *Mettl7a1*, is associated with the successful reprogramming trajectory; adding Mettl7a1 to the reprogramming cocktail increases the yield of induced endoderm progenitors. Together, these results demonstrate the utility of our lineage-tracing method for revealing the dynamics of direct reprogramming.

Direct lineage reprogramming bypasses pluripotency to convert cell identity between somatic states, yielding clinically valuable cell types[1]. However, these conversion strategies are generally inefficient, producing incompletely converted and developmentally immature cells that fail to fully recapitulate target cell identity[2,3]. The considerable heterogeneity that arises during reprogramming has hindered the study of the molecular mechanisms underlying lineage conversion. Single-cell RNA-sequencing analysis (scRNA-seq) has enabled fully converted cells to be distinguished from partially reprogrammed intermediates[4,5], although these analytical approaches typically result in the loss of spatial, temporal and lineage information. Elegant computational approaches can infer missing observations[6,7], but reconstruction of true reprogramming trajectories using these tools remains challenging. Although sophisticated lineage tracing solutions to connect cell history with fate are emerging, these protocols are either not compatible with high-throughput scRNA-seq[8–11], or require genome editing strategies that are not readily deployed in some systems[12–15].

To enable simultaneous single-cell profiling of cell identity and clonal history, we have developed 'CellTagging', a straightforward, high-throughput cell tracking method. Sequential lentiviral delivery of CellTags (heritable random barcodes) enables the construction of multi-level lineage trees. Here, we apply CellTagging to transcription factor-induced direct lineage reprogramming of mouse embryonic fibroblasts (MEFs) to induced endoderm progenitors (iEPs), a self-renewing cell type that has both hepatic and intestinal potential[3,16]. Generation of iEPs represents a prototypical cell fate engineering methodology, reflecting the inefficiency and infidelity of many reprogramming protocols[2,3]. CellTagging and tracking more than 100,000 cells during conversion to iEPs reveals two distinct trajectories: a route towards successfully reprogrammed cells, and an alternate path to a putative 'dead-end' state, marked by re-expression of fibroblast genes. Although few cells are successfully reprogrammed, clonally related cells tend to follow the same trajectories, suggesting that their reprogramming outcome may be determined from the earliest stages of lineage conversion. These clonal dynamics and lineages can be explored on our companion website, CellTag Viz (http://www.celltag.org/). In later stages of conversion, our analyses reveal expression of a putative methyltransferase, *Mettl7a1*, along the successful reprogramming trajectory. Adding this factor to the reprogramming cocktail increases the yield of successfully converted iEPs. Together, these findings demonstrate the utility of CellTagging for lineage reconstruction, providing molecular insights into reprogramming that serve to improve the outcome of this generally inefficient process.

Combinatorial indexing of cells to track clonal history

CellTagging is a lentivirus-based approach to uniquely label individual cells with heritable barcode combinations. CellTags are highly expressed and readily captured within each single-cell transcriptome, enabling recording of clonal history over time, in parallel with cell identity (Fig. 1a). Recovery of CellTag expression, followed by filtering and error correction, ensures sensitive and specific identification of clonally related cells (Extended Data Fig. 1a–g). The efficacy of this barcoding approach is demonstrated by CellTagging a 'species mix' of genetically distinct human 293T cells and MEFs (Extended Data Fig. 1h–j). This is further supported by labelling two independent biological replicates with the same CellTag library: whereas individual CellTags appear in both pools of cells, no combinatorial signatures of 2 or more CellTags are shared between replicates, confirming that clones are derived from distinctly labelled cells ($n = 4{,}141$ cells expressing 3.0000 ± 0.0004 (mean ± s.e.m.) unique CellTags per cell, Fig. 1b, c). Finally, CellTagging does not perturb cell physiology or reprogramming efficiency (Extended Data Fig. 2). Together, these data validate the utility of CellTagging to deliver unique, heritable labels into cells, permitting clonal relationships to be tracked longitudinally, with a high degree of confidence.

We next applied CellTagging to the direct reprogramming of fibroblasts to iEPs, driven by retroviral overexpression of the transcription factors FOXA1 and HNF4α (encoded by *Foxa1* and *Hnf4a*, respectively) in four independent biological replicates. To enable lineage reconstruction, we devised a sequential CellTagging scheme in which

[1]Department of Developmental Biology, Washington University School of Medicine in St Louis, St Louis, MO, USA. [2]Department of Genetics, Washington University School of Medicine in St Louis, St Louis, MO, USA. [3]Center of Regenerative Medicine, Washington University School of Medicine in St Louis, St Louis, MO, USA. [4]Present address: Nanomedicine Research Center, Department of Neurosurgery, Cedars–Sinai Medical Center, Los Angeles, CA, USA. *e-mail: s.morris@wustl.edu

© 2018 Springer Nature Limited. All rights reserved.

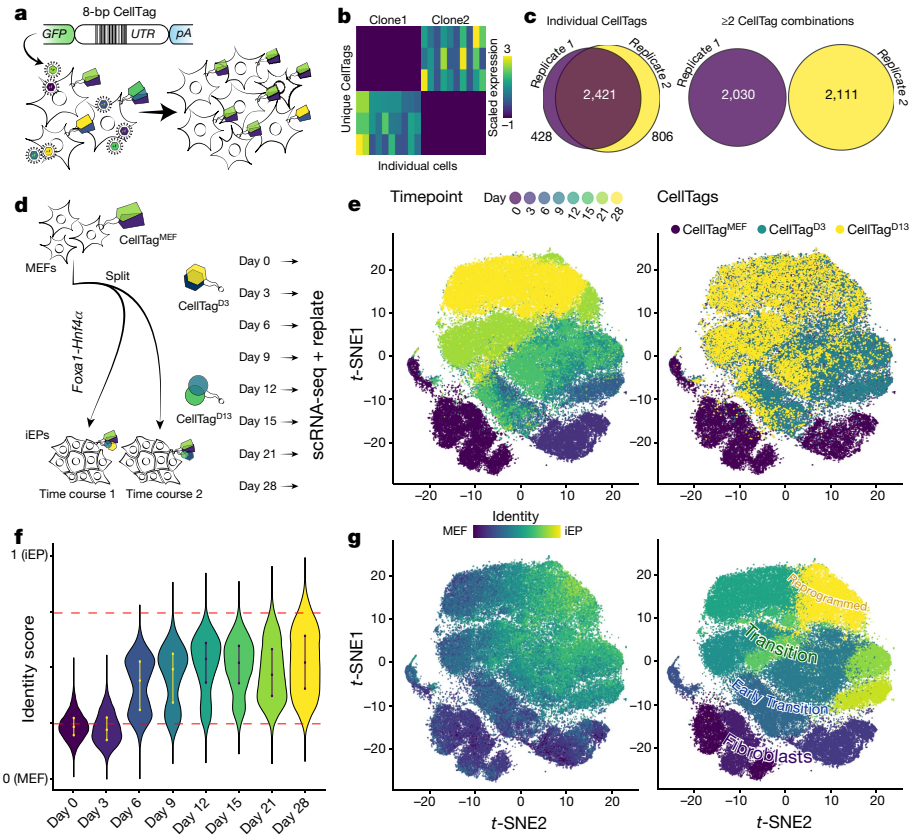


Fig. 1 | CellTagging: clonal tracking applied to reprogramming. a, The CellTagging workflow: a lentiviral construct contains an 8-bp random CellTag barcode in the 3′ untranslated region (UTR) of GFP, followed by an SV40 polyadenylation signal. Transduced cells express unique combinations of CellTags, resulting in distinct, heritable signatures, enabling tracking of clonally related cells. **b**, Representative CellTag expression in two clones, defined by unique combinations of three CellTags ($n = 10$ cells per clone). **c**, Left, overlap of individual CellTags in two independent biological replicates tagged with the same CellTag library. Right, CellTag signatures are not shared between the two replicates (replicate 1, $n = 8{,}535$ cells; replicate 2, $n = 11{,}997$ cells). **d**, Experimental approach: MEFs are tagged with the CellTag^{MEF} library, expanded for two days and then split for cell fate reprogramming in two independent biological replicates. Additional CellTagging was performed at 3 days (CellTag^{D3}) and 13 days (CellTag^{D13}) after initiation of reprogramming. Every 3–7 days, a sample of cells was collected for scRNA-seq, and the remaining cells were cultured. **e**, Visualization of scRNA-seq data. Projection of time points and CellTag expression onto a *t*-SNE plot (time courses 1 and 2, $n = 85{,}010$ cells). **f**, Scoring single-cell identity via quadratic programming. Cells scoring >0.75 (upper red line) are classified as iEPs; cells scoring <0.25 (lower red line) are classified as fibroblasts ($n = 85{,}010$ cells). **g**, Left, projection of identity scores onto the *t*-SNE plot. Right, designations of *t*-SNE clusters: fibroblast, early transition, transition and reprogrammed.

fibroblasts were transduced with an initial CellTag library, CellTag^{MEF}. Following a 48-h expansion period, these cells were split into independent biological replicates for reprogramming. Tagging with a second library (CellTag^{D3}) was performed at the end of the 3-day period of transcription factor delivery, followed by a third round (CellTag^{D13}) 13 days after the start of reprogramming, coinciding with the phenotypic emergence of iEPs. After sequencing, CellTags are assigned to rounds by demultiplexing on the basis of a short motif preceding the random CellTag region. Cells were collected every 3–7 days over the 28-day time course. A sample of cells from each time point was fixed in methanol for high-throughput droplet microfluidics-based scRNA-seq (Drop-seq[17] and 10x Genomics[18] platforms), and the remaining cells were replated to enable clonal growth and lineage reconstruction (Fig. 1d). In total, 104,887 single-cell transcriptomes were captured. Downstream analysis focused on data captured using the 10x Genomics platform (85,010 high-quality single-cell transcriptomes, merging time courses 1 and 2; Fig. 1e, Extended Data Fig. 3a–c, Supplementary Table 1). Canonical correlation analysis[19] demonstrates consistent replication across the sequencing technologies and biological replicates (Extended Data Fig. 3d, e).

Parallel capture of reprogramming and clonal dynamics

Using *t*-distributed stochastic neighbour embedding[6] (*t*-SNE), the 28-day reprogramming process resolves into 13 clusters of transcriptionally distinct cells (Extended Data Figs. 3f, g, 5a). CellTag expression is detected in 99% of cells, and CellTag^{MEF} expression is detected across all time points, CellTag^{D3} is detected from day 6, and CellTag^{D13} is detected from day 15 (Fig. 1e). Of 85,010 sequenced cells, 55,571 (65%) passed the threshold of at least two CellTags per cell that is required for tracking (Extended Data Fig. 4). To investigate dynamics of reprogramming, we first analysed gene expression for each cluster, revealing progressive silencing of fibroblast identity (Extended Data Fig. 5a, b, Supplementary Tables 2, 3). To track emergence of iEPs, we used quadratic programming[5] to score individual cell identities as a fraction of starting and target cell types, revealing that iEP identity is progressively gained from day 6 of reprogramming. Projection of identity scores onto the *t*-SNE plot localizes iEPs to cluster 2, coinciding with reprogramming days 21 and 28 (Fig. 1f, g). Further examination of this iEP-containing cluster identifies new markers, including apolipoprotein A1 (APOA1, encoded by *Apoa1*; Extended Data Fig. 5a, b, Supplementary Table 3). Immunostaining for APOA1 demonstrates protein-level co-expression with the canonical iEP marker E-cadherin (CDH1)[3,16] (Extended Data Fig. 5c–e). Although previous studies show that only around 1% of cells are successfully reprogrammed[3,16], we observe a high proportion of cells expressing *Apoa1*, beginning from day 6 ($62.5 \pm 5.5\%$; Extended Data Fig. 5b, d, e). Together, these observations suggest that many cells initiate reprogramming but few complete the transition to iEPs. Using expression of these markers,

© 2018 Springer Nature Limited. All rights reserved.

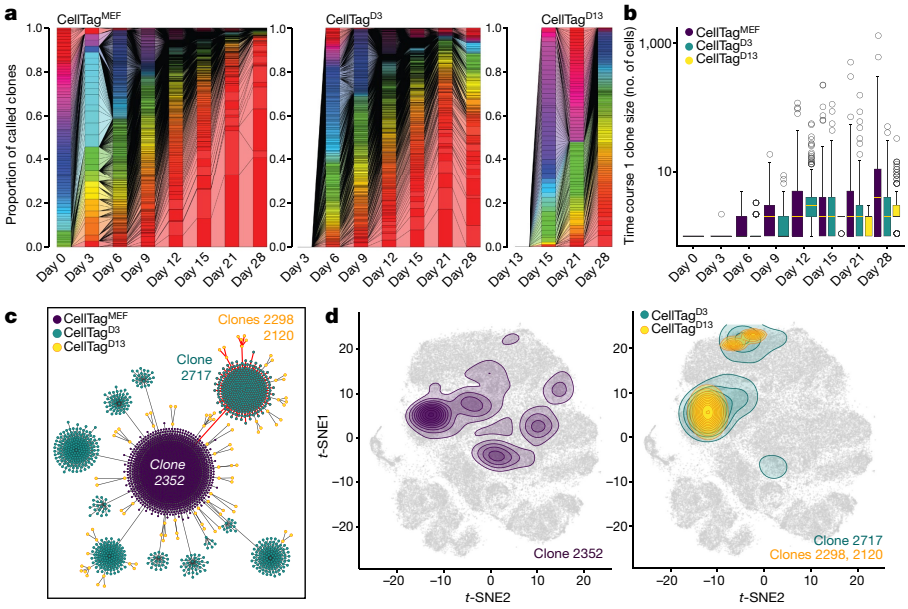


Fig. 2 | Tracking clonal dynamics of reprogramming and constructing lineage trees. a, Connected bar plots showing individual clones as a proportion of all clones during reprogramming, for each CellTagging round (time course 1, $n = 12{,}932$ cells, 1,031 clones). **b**, Mean number of cells per clone, per time point, for each round of CellTagging ($n = 1{,}031$ clones). **c**, Reconstruction and visualization of lineages using force-directed graph drawing. Each node represents an individual cell, and edges represent clonal relationships between cells: purple, CellTag^{MEF} clones; blue, CellTag^{D3} clones; yellow, CellTag^{D13} clones. **d**, Contour plots, representing cell density of each clone, projected onto the *t*-SNE for the lineage highlighted in red in **c** ($n = 2{,}199$ cells). All lineages and clone distributions can be explored with CellTag Viz (http://www.celltag.org/).

together with cell identity scores, we broadly partition the process into four phases: fibroblast, early transition, transition and reprogrammed (Fig. 1g, Extended Data Fig. 5b).

We next integrated clonal relationships into this single-cell landscape: from the 55,571 cells passing the threshold to support clone calling, we identified 27,020 cells possessing clonal relatives, on the basis of shared CellTag signatures. Defining a clone as three or more related cells, we identified 706 CellTag^{MEF} clones and 884 CellTag^{D3} clones. Because CellTag^{D13} clones had less time to expand, we also included related cell pairs for this later labelling, resulting in 561 clones (Supplementary Table 4). Consistent with the above validation experiments, examination of 10 major clones (defined as the ten largest clones based on number of cells) based on CellTag^{D3}-labelled replicates shows that the CellTag combinations used to identify clonally related cells were unique (Extended Data Fig. 6a). CellTags are reliably detected over a 10-week period; although their expression gradually diminishes over time, they are not silenced (Extended Data Figs. 4c, 6b–d). This demonstrates the advantage of our CellTag combinatorial indexing method for reliably labelling cells and tracking them over extended periods.

During reprogramming, we observed extensive clonal growth: CellTag^{MEF} clones reached an average size of 47 ± 22 cells per clone by day 28 (Fig. 2a, b, Extended Data Fig. 7a–d). Expanding at a similar rate, CellTag^{D3} clones were first detected from day 6, whereas smaller clones arose from CellTag^{D13}-labelling (Fig. 2a, b). In some instances, we observed rapid expansion of an individual clone during reprogramming (Extended Data Fig. 7d). This could not be reconciled with viral integration analysis (Supplementary Table 5), suggesting that the clonal growth we observed was associated with iEPs entering a self-renewing, progenitor-like state. As a consequence of this rapid expansion, iEPs were derived from only a small number of clones. We next sought to connect these clonal relationships over time, to trace the origins of successfully reprogramming cells. In this approach, we assume that the identity or state of each cell that we capture is representative of its collective clone. We find that gene expression is highly correlated among clonally related cells, suggesting that family members are likely to behave in a similar manner and share reprogramming outcomes (Extended Data Fig. 7e, f).

Lineage and reprogramming trajectory reconstruction

Sequential CellTagging enables the reconstruction of lineage trees and reprogramming trajectories. First, we apply force-directed graphing to construct hundreds of multi-level lineages (Extended Data Fig. 8a, b), which can be explored at http://www.celltag.org/. Figure 2c shows a representative lineage stemming from one CellTag^{MEF} clone, branching into CellTag^{D3} and CellTag^{D13} descendants. Next, to visualize the distribution of clonally related cells, we use contour plotting in combination with the *t*-SNE plot. This reveals considerable overlap of clones belonging to the same lineage, supporting our observation that clonally related cells are transcriptionally similar (Fig. 2d, Extended Data Fig. 8c, d). From these analyses, we observe enrichment or depletion of iEPs within many lineages. To quantify this, we re-clustered cells in the later stages of reprogramming, providing high-coverage clone information. Within this subset, 8% of cells are classified as fully reprogrammed iEPs (Fig. 3a; Extended Data Fig. 9a, b). We then performed randomized testing to identify major clones that were significantly enriched for or depleted of iEPs, yielding 20 iEP-enriched clones in which 20–50% of cells are fully reprogrammed. By contrast, we found 24 iEP-depleted clones in which less than 3% of cells are classified as iEPs (Fig. 3b).

iEP-enriched and iEP-depleted clones are clearly segregated on contour plots, suggesting the existence of discrete reprogramming trajectories; this is also supported by orthogonal pseudotemporal ordering analysis (Fig. 3c, d, Extended Data Fig. 9c, d). Quantification of these trajectories reveals a bifurcation at day 21, when successfully reprogramming clones transition through clusters 6 and 7, leading to the reprogrammed state at day 28. Conversely, these transition clusters are bypassed on the iEP-depleted trajectory, on which clones traverse cluster 4 on day 21, entering a putative reprogramming 'dead-end' by day 28 (Fig. 3e; Pearson's correlation coefficient, $r = -0.84$). To investigate the timing of the commitment to these trajectories, we quantified occupancy of CellTag^{D13}-labelled cells in reprogrammed and putative dead-end states (cluster1 and 3, respectively) at day 28. The distribution of clonally related cells between these states shows that they are restricted to one of the two states, indicating that reprogramming outcome is determined by day 13 (in 88 ± 8% of restricted clones; Extended Data Fig. 9e). These divergent routes appear to be rooted in distinct transcriptional states

© 2018 Springer Nature Limited. All rights reserved.

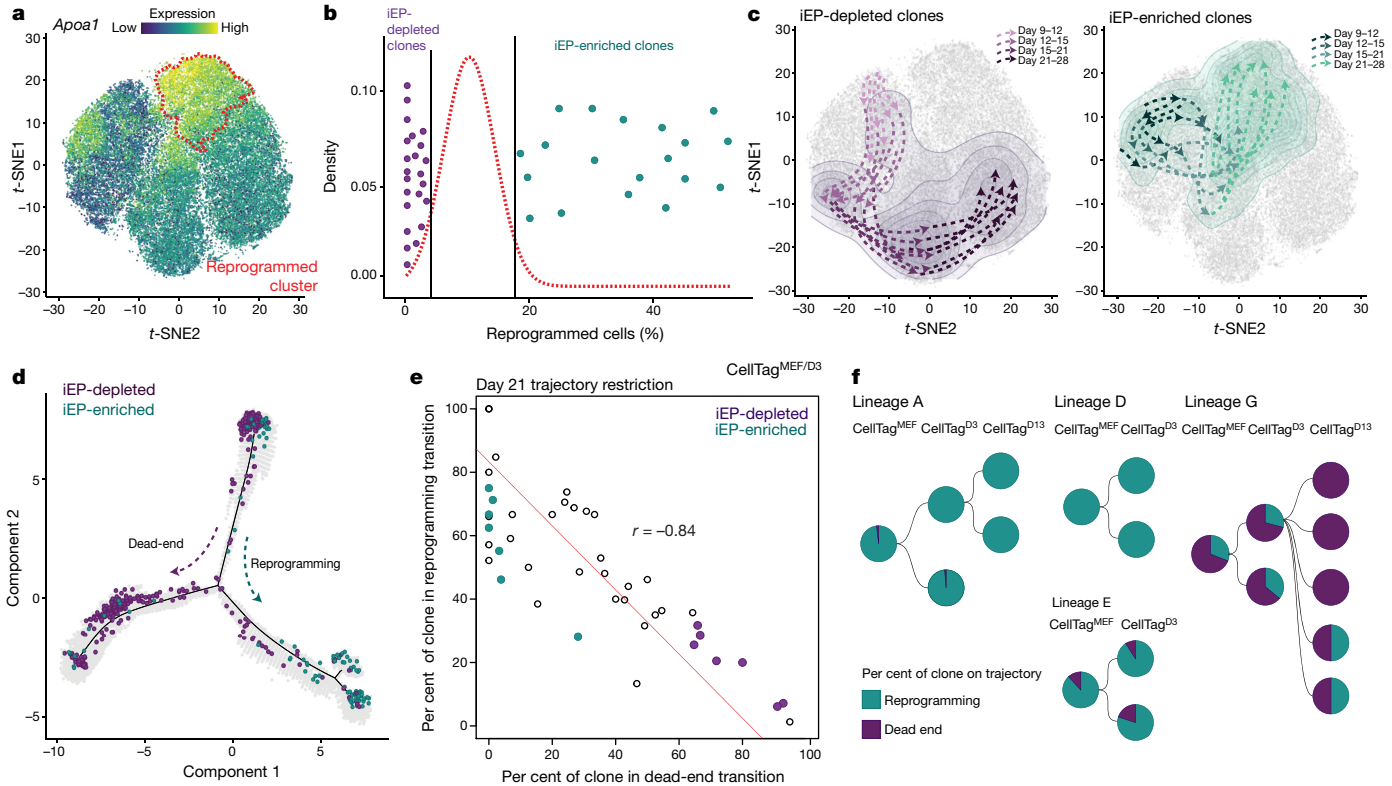


Fig. 3 | Mapping reprogramming trajectories and timing of cell fate commitment. a, *Apoa1* expression in a subset of cells from time courses 1 and 2 ($n = 48{,}515$ cells). Fully reprogrammed iEPs are outlined in red (cluster 1). **b**, Density plot of the mean proportion of reprogrammed cells for groups of randomly selected cells (defined by cluster 1 occupancy; $n = 59$ groups, 10,259 cells). Randomized testing of 59 CellTag$^{MEF/D3}$ clones (≥35 cells per clone, $n = 10{,}259$ cells) identifies iEP-enriched clones ($n = 20$ clones, 6,128 cells; $P < 0.05$) and iEP-depleted clones ($n = 24$ clones, 3,177 cells; $P < 0.05$). **c**, Clones spanning all time points were selected for further analysis. Trajectories showing connections between areas of highest clonal density across each day of reprogramming for iEP-depleted clones (left, $n = 7$ clones, 2,270 cells) and iEP-enriched clones (right, $n = 7$ clones, 1,037 cells). **d**, Pseudotemporal ordering of the time course 1 and 2 subset in **a**, with overlay of individual cells derived from iEP-enriched and iEP-depleted clones, defining reprogramming and dead-end trajectories ($n = 14$ clones, 3,307 cells). **e**, Proportions of clones occupying clusters 6 and 7 (reprogramming transition) or cluster 4 (dead-end transition) at reprogramming day 21 ($r = -0.84$, Pearson's correlation; $n = 44$ clones, 9,305 cells). **f**, Lineage trees of related clones, with the proportion of each clone contributing to reprogramming or dead-end trajectories shown ($n = 1{,}185$ cells).

as early as day 6 (Fig. 3c), suggesting that they are established early in the reprogramming process.

The existence of early-labelled clones that are biased in their reprogramming outcome, in addition to the shared transcriptional signatures that we observe between clonally related cells, suggests that cells do not reprogram in a stochastic manner. Here, sequential CellTagging and quantification of reprogramming outcome for each clone within a lineage allows us to probe the probability with which cells successfully generate iEPs. To study this, we identified lineages of CellTag^{D3}-labelled clones arising from common CellTag^{MEF}-labelled ancestors. For each clone within a lineage, we calculated the proportion of cells occupying reprogramming and dead-end trajectories. In a stochastic model of reprogramming, we would expect the post-reprogramming-induction, CellTag^{D3}-labelled clones from a common ancestor to follow different reprogramming trajectories. However, Fig. 3f shows that CellTag^{D3}-descendant clones reprogram with similar efficiencies to each other, and to their CellTag^{MEF}-labelled parent, particularly for those lineages reprogramming at high efficiency (Pearson's correlation coefficient, $r = 0.71$; Extended Data Fig. 9f). This suggests that reprogramming outcome may be determined at early stages. We considered the possibility that an 'elite' cell type that is predisposed to reprogram exists in the highly heterogeneous fibroblast starting population. To investigate this possibility, cells were first tagged and then split for reprogramming in two biological replicates. We identified 84 clones that appeared across both replicates; only 4 clones reprogrammed in both replicates (Supplementary Table 6), arguing against the existence of an elite reprogramming cell type in the fibroblast population.

Mettl7a1 delineates successful reprogramming

To investigate the molecular characteristics underpinning the distinct reprogramming paths, we compared cells between reprogramming and dead-end trajectories ($n = 2{,}074$ cells). Along the reprogramming trajectory, iEP identity scores gradually increase over time. By contrast, partial fibroblast identity is re-established with progression along the dead-end trajectory, supporting the suggestion that this represents a reprogramming impasse (Fig. 4a). Significant changes in gene expression between these two trajectories are apparent, including key elements of Wnt, Igf2 and HGF signalling pathways. The dead-end trajectory is enriched for imprinted gene expression (*Dlk1* and *Peg3*), in concert with reactivation of fibroblast gene expression and silencing of reprogramming transgenes. Many of these differences in gene expression are evident from day 6, including marked upregulation of *Apoa1* and concomitant downregulation of *Col1a2* on the reprogramming trajectory, supporting our observations that these outcomes are established from early stages. We did not detect significant differences in transgene expression between the two trajectories at these early stages, suggesting that transgene expression level is not a bifurcation driver (Fig. 4b, c, Extended Data Fig. 10a, b, Supplementary Table 7).

Focusing on later stages of reprogramming, we performed differential expression analysis of the trajectory bifurcation at day 21 (Supplementary Table 7). *Mettl7a1*, an as-yet-uncharacterized putative methyltransferase, was transiently and significantly upregulated along the successful reprogramming trajectory (Fig. 4b, c). METTL3, a related methyltransferase-like protein, catalyses N^6-methyladenosine (m^6A) modification of mRNA, and regulates stem-cell differentiation

© 2018 Springer Nature Limited. All rights reserved.

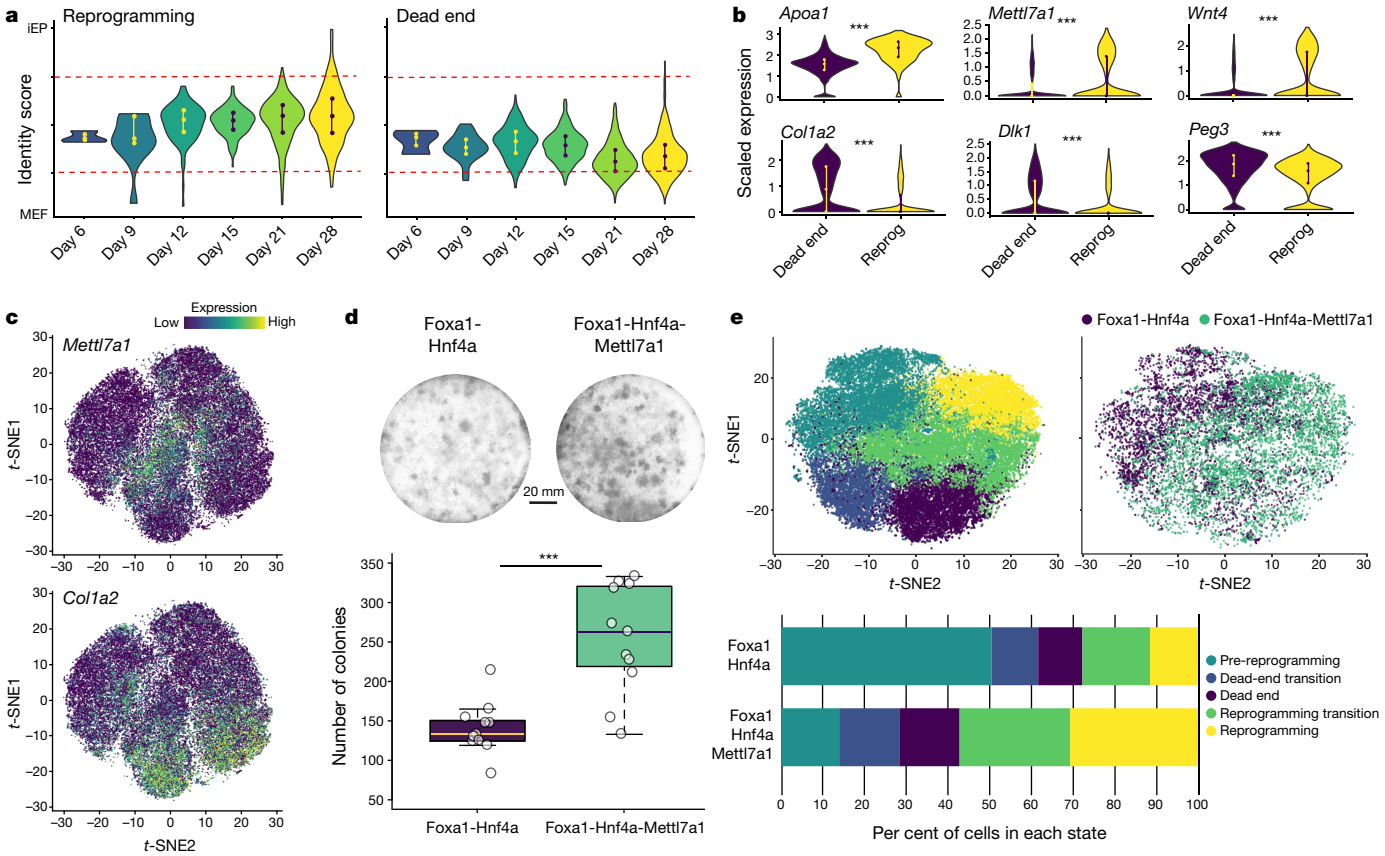


Fig. 4 | Molecular hallmarks of reprogramming trajectories. **a**, Identity scores of cells on the reprogramming (left, $n = 7$ clones, 1,037 cells) and dead-end trajectories (right, $n = 7$ clones, 1,037 cells, random downsampling from 2,270 cells) from reprogramming days 6 to 28. Cells scoring >0.75 (upper red line) are classified as iEPs, cells scoring <0.25 (lower red line) are classified as fibroblasts. **b**, Violin plots of significantly different ($P < 0.001$, permutation test, one-sided) gene expression between reprogramming and dead-end trajectories ($n = 14$ clones, 2,074 cells). **c**, Projection of *Mettl7a1* and *Col1a2* expression onto the *t*-SNE plot ($n = 48{,}515$ cells). **d**, Colony-formation assay (E-cadherin immunohistochemistry) for cells reprogrammed with Foxa1-Hnf4a or Foxa1-Hnf4a-Mettl7a1. Scale bar, 20 mm. Bottom, blinded and automated colony quantification. $n = 22$ technical replicates, 3 independent biological replicates; $P = 8 \times 10^{-5}$, one-sided *t*-test. **e**, Top, scRNA-seq analysis of 6,559 cells reprogrammed with Foxa1-Hnf4a and 6,559 cells (10,161 cells before random downsampling) reprogrammed with Foxa1-Hnf4a-Mettl7a1, 14 days after the start of reprogramming. Bottom, quantification of distribution of Foxa1-Hnf4a-Mettl7a1-reprogrammed cells across reprogramming stages, relative to that of Foxa1-Hnf4a-reprogrammed cells.

and reprogramming to pluripotency[20,21]. We therefore focused on *Mettl7a1* in the context of enhancing reprogramming efficiency. Addition of Mettl7a1 to the standard Foxa1-Hnf4a reprogramming cocktail resulted in a twofold increase in iEP colony formation (Fig. 4d). scRNA-seq of cells reprogrammed with Foxa1-Hnf4a or Foxa1-Hnf4a-Mettl7a1 reprogrammed cells shows that addition of Mettl7a1 to the reprogramming cocktail results in a threefold increase in the number of cells entering the fully reprogrammed state (Fig. 4e, Extended Data Fig. 10c–g). Inclusion of CellTags in these reprogramming experiments shows that under both control and Mettl7a1 conditions, the average number of cells per clone did not differ significantly between the two conditions (Extended Data Fig. 10h, i). Thus, Mettl7a1, rather than expanding existing iEPs, promotes a true increase in reprogramming efficiency.

Discussion

Here we have developed and validated a combinatorial indexing strategy, CellTagging, which enables simultaneous analysis of clonal history and cell identity at single-cell resolution. Our longitudinal dissection of Foxa1-Hnf4a-mediated direct lineage reprogramming to iEPs reveals two distinct conversion trajectories: one that leads to successful reprogramming, and one that leads to a dead-end state. We observe strong parallels between direct lineage reprogramming and induction of pluripotency: For instance, during induction of pluripotency, almost all cells initiate reprogramming, although transition to a fully pluripotent state is rare. This is characterized by two waves, or phases; in the second phase, a subset of cells are able to stably maintain the core pluripotency network[4,22]. In this context, the later bifurcation leading to the iEP state may parallel this second phase of reprogramming to pluripotency. Our identification of Mettl7a1 as a pro-reprogramming factor suggests that it may have an important role in the stabilization of iEP identity in later stages of lineage conversion.

Fibroblast-to-iEP conversion also shares a common feature with reprogramming to pluripotency with respect to inefficiency. On the basis of the low frequency of pluripotent cell generation, studies have suggested that the initiation and early phases of reprogramming are stochastic processes[4,23]. Our method of sequential CellTagging and lineage reconstruction enables reprogramming probabilities to be quantified. Tracking reprogramming outcome of clones derived from a shared ancestor strongly suggests that, in many cases, the trajectory of cell fate conversion is determined from the outset. If these early stages of reprogramming were stochastic, we would expect to see heterogeneity in reprogramming outcome between clones of the same lineage; however, we observe that clones of the same lineage follow similar reprogramming trajectories. Consistent with earlier studies[23], our CellTagging-and-split approach shows that clonally related cells—split into independent biological replicates—do not share reprogramming outcome, arguing against the existence of an elite cell type that is primed to reprogram. It is important to note here that, although we control the stoichiometry of the reprogramming factors, we do not control copy number or location of integration, which may produce a variable outcome between biological replicates.

© 2018 Springer Nature Limited. All rights reserved.

The evidence presented here suggests the existence of a privileged cell state in which reprogramming potential is predetermined. This is supported by several recent studies from reprogramming to pluripotency that also suggest the existence of a privileged state, or that cells can be coaxed into such a state via transient factor expression[24–28]. Furthermore, DNA barcode-based clonal analyses support a deterministic model of reprogramming[29]. Finally, scRNA-seq in combination with computational trajectory reconstruction suggests that reprogramming outcome can be predicted as early as two days following initiation via factor expression[30]. The next challenge will be to uncover the molecular hallmarks of this permissive state, enabling further improvements in reprogramming cells towards any desired cell identity with high efficiency and fidelity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0744-4.

Received: 3 June 2017; Accepted: 3 October 2018;
Published online 5 December 2018.

1. Vierbuchen, T. & Wernig, M. Direct lineage conversions: unnatural but useful? *Nat. Biotechnol.* **29**, 892–907 (2011).
2. Cahan, P. et al. CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903–915 (2014).
3. Morris, S. A. et al. Dissecting engineered cell types and enhancing cell fate conversion via *CellNet*. *Cell* **158**, 889–902 (2014).
4. Buganim, Y. et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
5. Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
6. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
7. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
8. Rodriguez-Fraticelli, A. E. et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).
9. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
10. Porter, S. N., Baker, L. C., Mittelman, D. & Porteus, M. H. Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines *in vitro* and *in vivo*. *Genome Biol.* **15**, R75 (2014).
11. Yao, Z. et al. A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* **20**, 120–134 (2017).
12. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
13. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
14. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
15. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
16. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**, 390–393 (2011).
17. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
18. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
19. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
20. Chen, T. et al. m^{6}A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* **16**, 289–301 (2015).
21. Batista, P. J. et al. m^{6}A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* **15**, 707–719 (2014).
22. Polo, J. M. et al. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617–1632 (2012).
23. Hanna, J. et al. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462**, 595–601 (2009).
24. Guo, S. et al. Nonstochastic reprogramming from a privileged somatic cell state. *Cell* **156**, 649–662 (2014).
25. Babos, K. N. et al. Balancing dynamic tradeoffs to drive cellular reprogramming. Preprint at https://www.biorxiv.org/content/early/2018/08/17/393934 (2018).
26. Rais, Y. et al. Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* **502**, 65–70 (2013).
27. Di Stefano, B. et al. C/EBPα poises B cells for rapid reprogramming into induced pluripotent stem cells. *Nature* **506**, 235–239 (2014).
28. Di Stefano, B. et al. C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. *Nat. Cell Biol.* **18**, 371–381 (2016).
29. Yunusova, A. M., Fishman, V. S., Vasiliev, G. V. & Battulin, N. R. Deterministic versus stochastic model of reprogramming: new evidence from cellular barcoding technique. *Open Biol.* **7**, (2017).
30. Schiebinger, G. et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. Preprint at https://www.biorxiv.org/content/early/2017/09/27/191056 (2017).

Acknowledgements We thank members of the Morris laboratory, and T. Druley and R. Mitra for critical discussions; S. McCarroll, E. Macosko and M. Goldman for advice establishing Drop-seq; B. Treutlein for quadratic programming assistance; J. Dick for the gift of the pSMAL backbone; K. Kniepkamp for help with CellTag Viz; and The Genome Technology Access Center in the Department of Genetics. This work was funded by National Institutes of Health (NIH) grants R01-GM126112, R21-HG009750; P30-DK052574; Silicon Valley Community Foundation, Chan Zuckerberg Initiative Grants HCA-A-1704-01646 and HCA2-A-1708-02799; The Children's Discovery Institute of Washington University and St. Louis Children's Hospital MI-II-2016-544. S.A.M. is supported by a Vallee Scholar Award; B.A.B.: NIH-T32HG000045-18; C.G.: NIH-5T32GM007200-42; S.E.W.: NIH-5T32GM007067-44; K.K.: Japan Society for the Promotion of Science Postdoctoral Fellowship.

Reviewer information *Nature* thanks L. Perié, M. Porteus, L. Vallier and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.A.B. and S.A.M. conceived the research. S.A.M. led experimental work, assisted by B.A.B., W.K., C.G., S.E.W. and T.S. B.A.B. and W.K. led computational analysis, assisted by K.K. and supervised by S.A.M. All authors participated in interpretation of data and writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0744-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0744-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.A.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. Except where stated, the investigators were not blinded to allocation during experiments and outcome.

Mice and derivation of mouse embryonic fibroblasts. MEFs were derived from embryonic day (E)13.5 C57BL/6J embryos. (The Jackson laboratory: 000664). Heads and visceral organs were removed and the remaining tissue was minced with a razor blade and then dissociated in a mixture of 0.05% trypsin and 0.25% collagenase IV (Life Technologies) at 37 °C for 15 min. After passing the cell slurry through a 70-μM filter to remove debris, cells were washed and then plated on 0.1% gelatin-coated plates, in DMEM supplemented with 10% FBS (Gibco), 2 mM L-glutamine and 50 mM β-mercaptoethanol (Life Technologies). All animal procedures were based on animal care guidelines approved by the Institutional Animal Care and Use Committee.

Lenti- and retrovirus production. Lentiviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the pSMAL-CellTag construct (see below), along with packaging constructs pCMV-dR8.2 dvpr (Addgene plasmid 8455), and pCMV-VSVG (Addgene plasmid 8454). Constructs were titred by serial dilution on 293T cells. Hnf4a-T2A-Foxa1 and Mettl7a1 were cloned into the pGCDN-Sam retroviral construct and packaged with pCL-Eco (Novus Biologicals, NBP2-29540), titred on fibroblasts. We opted to generate a bicistronic Hnf4a-Foxa1 construct, based on the T2A sequence to increase the consistency of reprogramming via maintenance of exogenous transcription factor stoichiometry. Virus was collected 48 h and 72 h after transfection and applied to cells immediately following filtering through a low-protein binding 0.45-μm filter.

CellTagging methodology. To generate CellTags, we introduced an 8-bp variable region into the 3′UTR of GFP in the pSMAL lentiviral construct[31], using a gBlock gene fragment (Integrated DNA Technologies) and megaprimer insertion. This approach relies on the presence of 60-bp 'arms' in the gene fragment that are homologous to the desired plasmid insertion site. The fragments were then introduced into the plasmid using PCR, followed by DpnI (New England Biolabs) treatment to digest non-modified plasmid. All the recovered DNA from bacterial transformation (Stellar Competent Cells, Takara Biosciences) was grown overnight in liquid culture, followed by maxi-prep extraction of the plasmid DNA. This complex library of CellTag constructs was used to generate lentivirus (above) which was then used to transduce fibroblasts at a multiplicity of infection of ~3–4. For CellTag versions 2 and 3, a short 6-bp sequence was also included, just upstream of the variable CellTag region. For CellTag version 2, this sequence motif is GTGATG. For CellTag version 3, this sequence motif is TGTACG. For both Drop-seq and 10x Genomics-based experiments, the starting fibroblast population was transduced with CellTag version 1 (denoted as $\text{CellTag}^{\text{MEF}}$) for 24 h, followed by washing and culture for a further 48 h. At this point, cells were split, with one portion taken for Drop-seq/10x Genomics and two portions replated for reprogramming to iEPs in two biological replicates. For 10x Genomics-based experiments, cells were tagged again, immediately following 72 h of reprogramming, with CellTag version 2 (denoted as $\text{CellTag}^{\text{D3}}$). One further round of CellTagging followed at day 13 post-initiation of reprogramming with CellTag version 3 (denoted as $\text{CellTag}^{\text{D13}}$). Pooled CellTag libraries have been deposited at Addgene: https://www.addgene.org/pooled-library/morris-lab-celltag/, pSMAL-CellTag-V1 (pooled library #115643); pSMAL-CellTag-V2 (pooled library #115644); pSMAL-CellTag-V3 (pooled library #115645).

Generation and collection of iEPs. Early passage MEFs (<passage 6) were reprogrammed with modifications to the described protocols[16]. We modified this protocol, transducing cells every 12 hours for 3 days, with fresh Hnf4a-T2A-Foxa1 retrovirus in the presence of 4 μg/ml protamine sulfate (Sigma-Aldrich). These transduced cells were then cultured on 0.1% gelatin-treated plates for 1 week in hepato-medium (DMEM:F-12, supplemented with 10% FBS, 1 μg/ml insulin (Sigma-Aldrich), 100 nM dexamethasone (Sigma-Aldrich), 10 mM nicotinamide (Sigma-Aldrich), 2 mM L-glutamine, 50 mM β-mercaptoethanol (Life Technologies), and penicillin–streptomycin, containing 20 ng/ml epidermal growth factor (Sigma-Aldrich)). After 7 days of culture, the cells were transferred onto plates coated with 5 $\mu g/cm^2$ Type I rat collagen (Gibco, A1048301). For Drop-seq based experiments (two independent biological replicates), with a cell capture rate of 5%, 2×10^5 cells were initially seeded, and cells were collected every 7 days. At each collection, cells were gently dissociated in TrypLE Express (Gibco), and 1.5×10^5 cells were collected for Drop-seq, replating and culturing the remaining cells. For 10x Genomics-based experiments, with a cell encapsulation rate of up to 60%, 5×10^4 cells were initially seeded and collected every 3–7 days. At each cell collection, 3×10^4 dissociated cells were fixed in methanol, and the remaining cells were replated and cultured. Methanol fixation was performed as previously described[32]. In brief, cells were collected and washed in phosphate buffered saline (PBS), followed by resuspension in ice-cold 80% methanol in PBS, with gentle vortexing. These cells were stored at −80 °C for up to three months, and processed in the same batch on the 10x Genomics platform (below). iEP lines at the end of reprogramming tested negative for mycoplasma.

Immunostaining. iEP cells were grown in 4-Chamber Culture Slides (Falcon #354114) and fixed in 4% paraformaldehyde. Cells were permeabilized in 0.1% Triton-X100, followed by blocking in 10% fetal bovine serum in PBS (blocking buffer). Primary antibody, goat apolipoprotein A-I antibody (1:100, Novus Biologicals, NB600-609, lot: 30506) or mouse E-cadherin antibody (1:50, BD Biosciences, 610181, Clone: 36/E-cadherin, lot: 7187865) in blocking buffer was applied overnight before washing and applying secondary antibody: Alexa Fluor 555 rabbit anti-goat IgG (1:1000, Invitrogen A-21431) or Alexa Fluor 488 goat anti-mouse IgG (1:1000, Invitrogen A-32723), diluted in blocking buffer. Nuclear staining was performed with 300 nM DAPI in PBS. Slides were mounted with ProLong Gold antifade reagent (Invitrogen P36930). Images were captured using a Zeiss Axio Imager Z2 fluorescent microscope.

Mettl7a1 reprogramming and colony formation assay. Mouse *Mettl7a1* (NM 027334, Origene: MC205948) was sub-cloned into the retroviral vector, pGCDN-Sam[16], and retrovirus was produced as described above. For comparative reprogramming experiments, MEFs (1.2×10^5 cells per 6-cm plate, in 3 independent biological replicates) were serially transduced over 72 h (as above), followed by splitting and seeding at 4×10^4 cells per well of a 6-well plate to generate technical replicates. In control experiments, virus produced from an empty vector control expressing only GFP was added to the Foxa1-Hnf4a reprogramming cocktail. In Mettl7a1 experiments, virus produced from the Mettl7a1-IRES-GFP construct was added to virus containing Hnf4a and Foxa1. *Mettl7a1* overexpression was confirmed by preparing RNA from cells transduced with Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1 using the RNeasy kit (Qiagen). Following cDNA synthesis (Maxima cDNA synthesis kit, Life Tech), quantitative reverse transcription with PCR (qRT–PCR) was performed to quantify *Mettl7a1* overexpression (TaqMan Probe: Mm03031185_sH, TaqMan qPCR Mastermix, Applied Biosystems). Cells were reprogrammed for two weeks, at which point the cells in some wells were dissociated and fixed in methanol for 10x Genomics-based single-cell analysis (details below). The remaining wells were processed for colony-formation assays: cells were fixed on the plate with 4% paraformaldehyde, permeabilized in 0.1% Triton-X100 then blocked with Mouse on Mouse Elite Peroxidase Kit (Vector PK-2200). Mouse E-cadherin antibody (1:100, BD Biosciences) was applied for 30 min before washing and processing with the VECTOR VIP Peroxidase Substrate Kit (Vector SK-4600). Colonies were visualized on a flatbed scanner, adding heavy cream to each well to increase image contrast. Colonies were counted, using the colony counter ImageJ plugin (https://imagej.nih.gov/ij/plugins/colony-counter.html). These analyses were blinded.

Drop-seq. Cells were dissociated using TrypLE Express (Gibco), washed in PBS containing 0.01% BSA and diluted to 100 cells/μl, then processed by Drop-seq within 15 min of their collection. Drop-seq was performed as previously described[17] (http://mccarrolllab.com/dropseq/). In brief, cells and beads were diluted to an estimated co-occupancy rate of 5% upon co-encapsulation: 1×10^5 cells/ml and 1.2×10^5 beads/ml. Two independent lots of beads (Macosko-2011-10, ChemGenes) were used: 091615 (time course 3) and 032516B (time course 4). Emulsions were collected and broken using 1 ml of Perfluorooctanol (Sigma) for 15 ml of emulsion, followed by washing in 6× saline-sodium citrate (SSC) buffer to recover beads. Reverse transcription was then performed using the Maxima H Minus Reverse Transcriptase kit (EP0752, Life Tech). After treatment with 2,000 U/ml of ExonucleaseI (New England Biolabs), aliquots of 2,000 beads (representing ~100 single-cell transcriptomes for a cell-bead co-encapsulation rate of 5%) were amplified by PCR for 13 cycles, using Kapa HiFi Hotstart Readymix (Kapa Biosystems). The PCR product resulting from this reaction was purified by addition of 0.6× AMPure XP beads (Beckman Coulter). Six hundred picograms of this purified cDNA product from an estimated 5,000 cells was tagmented using Nextera XT according to the manufacturer's protocol (Illumina). The resulting cDNA library was again purified using 0.6× AMPure XP beads, followed by 1× AMPure XP beads. cDNA concentrations were assessed by Tapestation (Agilent) analysis. Libraries were sequenced on an Illumina HiSeq 2500, with custom priming (Read1CustSeqB Drop-seq primer).

10x Genomics procedure. For single-cell library preparation on the 10x Genomics platform, we used: the Chromium Single Cell 3′ Library and Gel Bead Kit v2 (PN-120237), Chromium Single Cell 3′ Chip kit v2 (PN-120236) and Chromium i7 Multiplex Kit (PN-120262), according to the manufacturer's instructions in the Chromium Single Cell 3′ Reagents Kits V2 User Guide. Just before cell capture, methanol-fixed cells were placed on ice, spun at 3,000 r.p.m. for 5 min at 4 °C, followed by resuspension and rehydration in PBS, according to a previously described method[32]. Seventeen thousand cells were loaded per lane of the chip, aiming for capture of 10,000 single-cell transcriptomes. All samples were processed in parallel, on the same day. Resulting cDNA libraries were quantified on an Agilent Tapestation and sequenced on an Illumina HiSeq 3000.

Viral integration analysis. Genomic DNA was prepared from control MEFs and iEPs derived from clone 1 (time course 4), using the DNeasy Blood & Tissue kit (Qiagen). Sample quality was assessed by Qubit DNA Assay Kit and gel electro-

© 2018 Springer Nature Limited. All rights reserved.

phoresis. Library construction was carried out using the Nextera XT Library prep kit (Illumina) following the manufacturer's recommendations. The lentivirus integration boundary sequence was enriched by amplification using primers specific for lentivirus long terminal repeat (LTR) and the Nextera XT adaptor sequence. Two separate PCR reactions were performed for each sample, one for 3′ LTR and another for 5′ LTR. The final PCR was performed to add Illumina sequencing adapters with unique barcodes for each sample. The libraries for each sample were pooled into a final library and assessed by Qubit DNA assay, Agilent Bioanalyzer and qRT–PCR. The library was sequenced on the NextSeq 500 system using the 150 Cycle High Output flow cell. Fastq data was extracted from the NextSeq system using bcl2fastq and the quality control of the data was performed using FastQC. Fastq reads were aligned to the mouse reference genome (GRCm38) using BWA MEM. De-duplication was performed using Samtools. Peak calling and comparison between two samples for putative lentivirus integration site was performed using MACS2.

Library preparation and sequencing of CellTag plasmid libraries for whitelist generation. Library construction was carried out using the Nextera XT Library prep kit (Illumina), following the manufacturer's recommendations. The CellTag region was enriched by amplification using primers specific for the pSMAL lentivirus GFP UTR and the Nextera XT adaptor sequence. A final PCR was performed to add Illumina sequencing adapters. The libraries for each CellTag version were pooled and assessed by Tapestation (Agilent). The library was sequenced on an Illumina MiSeq. Reads that contained the CellTag motif were identified (see 'CellTag demultiplexing'). A 90% percentile cut-off in terms of reads reported for each CellTag was used to select CellTags for inclusion on the whitelist of cell barcodes.

10x Genomics and Drop-seq alignment, digital gene expression matrix generation. The Cell Ranger v.2.1.0 pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to process data generated using the 10x Chromium platform. This pipeline was used in conjunction with a custom reference genome, created by concatenating the sequences corresponding to the Hnf4a-T2A-Foxa1 transgene and the GFP-CellTag transgene as new chromosomes to the mm10 genome. The unique UTRs in the Hnf4a-T2A-Foxa1 and GFP-CellTag transgene constructs allowed us to monitor transgene expression. To create Cell Ranger-compatible reference genomes, the references were rebuilt according to instructions from 10x Genomics (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/references). To achieve this, we first created a custom gene transfer format (GTF) file, containing our transgenes, followed by indexing of the FASTA and GTF files, using Cell Ranger mkgtf and mkref functions. Following this step, the default Cell Ranger pipeline was implemented, with the filtered output data used for downstream analyses. For Drop-seq analysis, raw reads were processed, filtered, and aligned as previously described[17], including correction of barcode synthesis errors. This process and the required tools, are further outlined online in the Drop-seq Alignment Cookbook (http://mccarrolllab.com/dropseq/). To facilitate downstream analyses the reference genome used during alignment was modified to include the transgenic sequences above. Processed reads were aligned to a custom genome build, using STAR. Across all experiments, the mean number of confidently mapped reads per cell was 38,259 (Supplementary Table 1).

Following alignment, digital gene expression (DGE) matrices were generated for each time point, for all time courses. Drop-seq DGEs were aggregated using a custom R script. Merged 10x Genomics DGE files were generated using the aggregation function of the Cell Ranger pipeline. We then performed initial filtering of these DGE files as a quality control step. We first removed cells with a low number (<200) of unique detected genes. We then removed cells for which the total number of unique molecular identifiers (UMIs) (after log transformation) was not within three standard deviations of the mean. This was followed by the removal of outlying cells with an unusually high or low number of UMIs given their number of reads by fitting a loess curve (span = 0.5, degree = 2) to the number of UMIs with number of reads as predictor (after log transformation), removing cells with a residual more than three standard deviations away from the mean. This process was also used to remove cells for with unusually high or low number of genes given their number of UMIs. Finally, we removed cells in which the proportion of the UMI count attributable to mitochondrial genes was greater than 10% (for Drop-seq-based experiments) or 20% (for 10x Genomics-based experiments).

Data normalization and scoring of cell cycle phase. Following DGE filtering, cell cycle scores were generated for each cell and data were normalized. Cell cycle scores were generated using a pre-defined classifier to assign cell cycle phase for each cell. This classifier was built from training data by identifying pairs of genes where the difference in expression within each pair changed sign across phases. Cell cycle phase was assigned to each cell by examination of the sign of the difference in test data. After calculating the cell cycle scores, the data was normalized using the 'deconvolution' method. This method pools cells and combines the expression values of the cells in a pool. The pooled expression values are used to calculate size-factors for normalization. These pool-based normalization factors can then be deconvoluted into cell-specific normalization factors, which are then used to normalize the expression of each cell. This deconvolution normalization method is an attempt to address the abundance of zero counts that is prevalent to scRNA-seq. The cell cycle scores and data normalization was facilitated by the Scater package[33], available on Bioconductor.

CellTag demultiplexing. Reads containing the CellTag sequence were extracted from the processed and filtered BAM files produced by the 10x Genomics and Drop-seq pipelines. Reads that contained the CellTag motif were identified from the following sequences: CellTagV1 ($CellTag^{MEF}$): CCGGTNNNNNNNNGAATTC, CellTagV2 ($CellTag^{D3}$): GTGATGNNNNNNNNGAATTC, CellTagV3 ($CellTag^{D13}$): TGTACGNNNNNNNNGAATTC. Following extraction of reads from the BAM file, a custom gawk script was used to parse the output, capturing the read ID, sequence, cell barcode, UMI, CellTag sequence and aligned genes for each read. This parsed output was then used to construct a cell barcode × CellTag UMI matrix. CellTags were grouped by cell barcodes and then the number of unique UMIs for each cell barcode–CellTag pair was counted. The matrix was then filtered to remove any cell barcodes not found in the filtered Cell Ranger and Drop-seq output files. Finally, the CellTags were filtered to remove any that were represented by $\leq$1 UMI. The construction and filtering of the CellTag UMI matrix was accomplished using a custom R script. Using this matrix, an error-correction step was then performed to amend PCR and sequencing errors: CellTags one edit-distance apart were collapsed on a cell-by-cell basis, using Starcode[34], an algorithm to determine which sequence pairs lie within a given Levenshtein distance, merging matched pairs into clusters of similar sequences. This filtered CellTag UMI count matrix was then used for all downstream clone and lineage analysis.

CellTag filtering and clone calling. The CellTag matrix was initially filtered by removing CellTags that do not appear on the whitelists generated for each CellTag plasmid library (see 'Library preparation and sequencing of CellTag plasmid libraries for whitelist generation'). CellTags appearing in $>$5% of cells in the first time point were also removed as this would suggest dominance of the library by individual CellTags that would interfere with accurate clone-calling. The requirement for this filtering was rare. Cells expressing more than 20 CellTags (likely to correspond to cell multiplets), and less than 2 CellTags per cell were filtered out. To identify clonally related cells, Jaccard analysis using the R package Proxy was used to calculate the similarity of CellTag signatures between cells. A Jaccard score of $>$0.7 was used as a cut-off to identify cells highly likely to be related, on the basis of our experimental findings. We found this cut-off to be stringent enough for unrelated cells not to be connected, but in a small number of instances, we found related cells that were not connected, probably owing to CellTag errors that were not corrected, or CellTag dropout. These related cells were united as part of lineage construction, below. Clones were defined as groups of 3 or more related cells (for $CellTag^{MEF}$, $CellTag^{D3}$), or 2 or more related cells (for $CellTag^{D13}$) identified using a custom R script. Clones were visualized using the Corrplot package with hierarchical clustering, contour plotting using ggplot2, or using force-directed network graphs (see below). Clones were called on cells pre-filtered for numbers of genes, UMIs and mitochondrial RNA content.

Seurat, Monocle and quadratic programming analyses. After filtering and normalization, the R package Seurat[6] was used to cluster and visualize cells. As the data were already normalized, they were loaded into Seurat without normalization, scaling or centring. Along with the expression data, metadata for each cell was collected, including information such as clone identity, cell cycle phase, and time point (Supplementary Table 4). Seurat was used to remove unwanted variation, regressing out number of UMIs, proportion of mitochondrial UMIs and cell cycle scores. Next, highly variable genes were identified and used as input for dimensionality reduction via principal component analysis (PCA). The resulting PCs and the correlated genes were examined to determine the number of components to include in downstream analysis. These PCs were then used as input to cluster the cells, visualizing these clusters using *t*-SNE. Semi-supervised Monocle[7] analysis was used to order cells in pseudotime, based on expression of the fibroblast marker *Col1a2* and the iEP marker *Apoa1*. Quadratic programming[5] was used to score fibroblast and iEP identity. This approach was modified to use bulk expression data of MEF and iEP collected previously[16] and whole transcriptome profiles of the two cell types were used for identity score calculation. The R package QuadProg was used for quadratic programming to generate cell identity scores. Investigators were blinded to allocation in the orthogonal pseudotemporal ordering analysis.

Lineage visualization via construction of force-directed network graphs. Network graphs were constructed by integrating all data for all rounds of CellTagging. In the graphs, each node represents an individual cell, and edges represent clonal relationships between cells. First, using a custom R-based script, cells were assembled into sub-clusters, according to $CellTag^{MEF}$, $CellTag^{D3}$, and $CellTag^{D13}$ information. Then, these sub-clusters were connected to each other to build lineages of related cells, connected across the different rounds of CellTagging—that is, two different $CellTag^{D3}$ clones sharing the same $CellTag^{MEF}$

© 2018 Springer Nature Limited. All rights reserved.

labels are part of the same lineage. Using this approach, we identified collisions in $4.5 \pm 1.1\%$ of clones—a collision is defined as one clone sharing two or more parents. In these cases, we inspected the CellTag signature for each clone and united any clones that had been split, reducing the collision rate to $0.9 \pm 0.6\%$. The resulting networks were visualized as force-directed network graphs using Cytoscape 3.6.0 and Allegro Layout. Allegro spring-electric was used as the layout protocol to render force-directed network graphs. Individual graphs for each lineage can be explored with our Shiny-based interactive platform, CellTag Viz (http://www.celltag.org/).

Trajectory discovery by randomized testing. To identify clones with an enriched or depleted rate of iEP generation, we used randomized testing to evaluate whether each clone (of at least 35 cells in size) possesses a similar percentage of fully reprogrammed cells relative to a randomly selected population of the same size. Here, the percentage of reprogrammed cells is defined as the proportion of cells within each group found in the reprogrammed cluster, as defined by Seurat. Two groups, cells of the clone and that of the overall population, are compared with the null percentage calculated using the cells in each clone. Let N represent the number of cells in each clone and M represent the remaining cell population size. We pool the two groups of cells (size $= N + M$) and resample N random cells, without replacement, from the pooled cells $(N + M)/N$ times such that every possible separation with ending groups of size N and M can be sampled and captured. During this process, the percentage is calculated based on the N randomly sampled cells. With the percentage calculated, P values can be evaluated based on the proportion of randomly sampled cells with a percentage greater than or equal to the null percentage. Using the P value of <0.05 (>0.95 for the other tail), we identified clones that were enriched or depleted for reprogrammed cells. These calculations were performed using a custom R-based script. Clones with at least 35 cells were selected to increase the statistical power of this analysis. For permutation testing to analyse differences in trajectory-specific gene expression, a custom Python-based script was used.

Reagent and protocol availability. Pooled CellTag libraries have been deposited and are available from Addgene: https://www.addgene.org/pooled-library/morris-lab-celltag/. A working protocol can be accessed via protocols.io https://doi.org/10.17504/protocols.io.vawe2fe.

Code availability. Code for processing of CellTag data, clone-calling, and construction of lineage trees is available on GitHub (https://github.com/morris-lab).

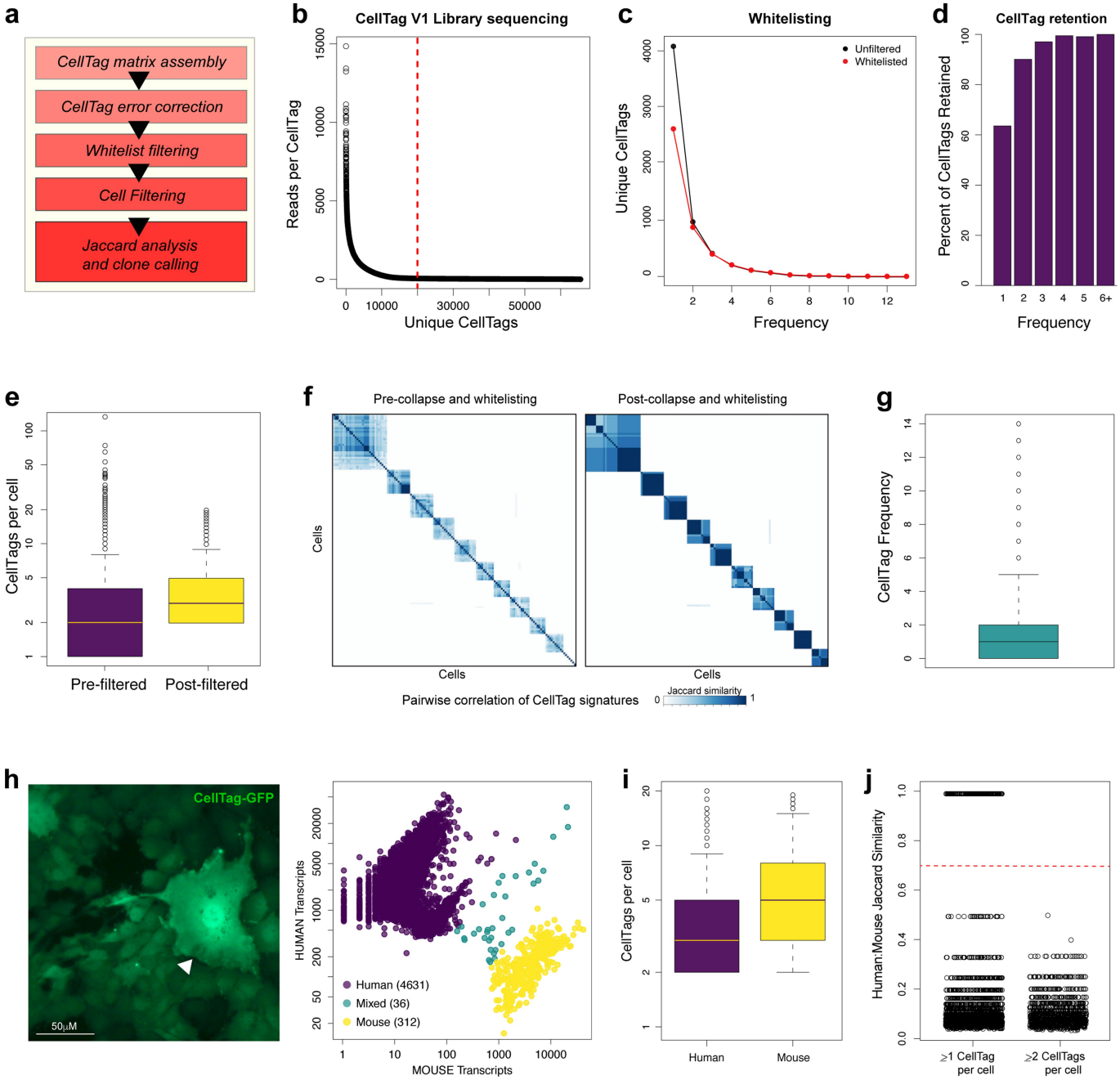
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All source data, including sequencing reads and single-cell expression matrices, are available from the Gene Expression Omnibus (GEO) under accession code GSE99915.

31. van Galen, P. et al. The unfolded protein response governs integrity of the haematopoietic stem-cell pool during stress. *Nature* **510**, 268–272 (2014).
32. Alles, J. et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* **15**, 44 (2017).
33. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **347**, btw777 (2017).
34. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913–1919 (2015).

© 2018 Springer Nature Limited. All rights reserved.

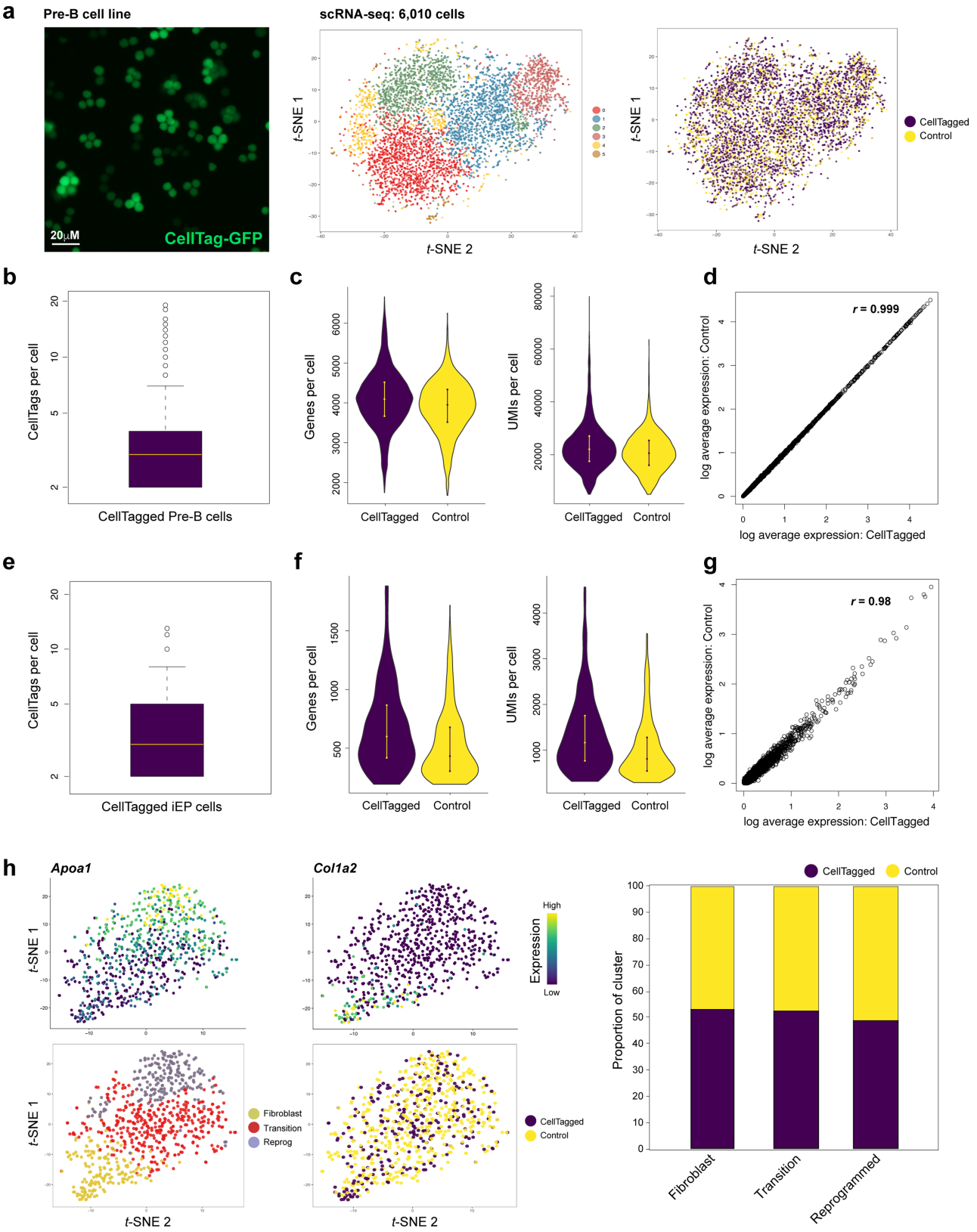


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 1 | CellTag processing and species-mixing validations. a, Schematic of the CellTag processing and filtering pipeline: CellTag sequences are first extracted from aligned sequencing reads, followed by construction of a matrix of CellTag expression in each cell. To mitigate potential artefacts arising as a result of PCR and sequencing errors, we implemented an error-correction step, collapsing similar barcodes one edit-distance apart, on a cell-by-cell basis. An initial filtering step then removes any CellTags that do not appear on a whitelist of CellTags that are confirmed to exist in the complex lentiviral library. A second filtering step removes cells expressing less than two or more than 20 unique CellTags. Using this filtered dataset, Jaccard analysis is then applied (using the R package, Proxy) to identify related cells, based on CellTag signature similarity, allowing clones to be called. **b**, Generation of the CellTag whitelist. Following CellTag lentiviral plasmid sequencing, CellTags were extracted from the raw fastq files via identification of the adjacent motifs as described in Methods (see Methods, 'CellTag demultiplexing'). A 90th percentile cut-off in terms of reads reporting each CellTag was used to select CellTags for inclusion on the whitelist. Of a possible 65,536 unique combinations, we detected 19,973 sequences passing this 90th percentile of read counts. Data for CellTag version 1 ($CellTag^{MEF}$) is shown here. Whitelist creation was also performed for CellTag versions 2 ($CellTag^{D3}$) and 3 ($CellTag^{D13}$). **c**, **d**, CellTag frequency (**c**), that is, how many times each CellTag is detected in a population of transduced cells, before (black) and after (red) removal of CellTags that do not feature on the whitelist. This whitelisting predominantly results in the removal of CellTags that appear only once; singletons that are likely to arise owing to sequencing and PCR errors. This is reflected in the histogram in **d**, showing that only 60% of singleton CellTags detected are retained, whereas over 90% of CellTags appearing in two or more cells are retained. **e**, Mean CellTags per cell pre- and post-CellTag pipeline filtering. Cells in this figure correspond to the cells shown in Fig. 1b, c (replicate 1: $n = 8{,}535$ cells; replicate 2: $n = 11{,}997$ cells). **f**, Pairwise correlation scores (Jaccard similarity) and hierarchical clustering of 10 major clones arising from this tag and trace experiment. Hierarchical clustering is based on each cell's Jaccard correlation relationships with other cells, where each defined 'block' of cells represents a clone. Left, scoring and clustering of pairwise correlations, before whitelisting and filtering. Right, after whitelisting and filtering, pairwise correlations are stronger and more cells are detected within each clone ($n = 869$ cells). **g**, CellTag frequency metric: each detected CellTag appears in less than two cells ($n = 9{,}072$ cells in total) at the start of the experiment, on average. The library is therefore not dominated by any abundant CellTags, which would potentially generate false-positive results. **h**, A species mixing experiment, consisting of a mixture of human 293T cells and MEFs (left), labelled with ~3–5 CellTags per cell and expressing GFP as a result. A fibroblast (white arrow) is visible within a colony of 293T cells. Scale bar, 50μ M. Seventy-two hours after transduction, cells were collected and processed for Drop-seq. Right, following sequencing and alignment, cells were assigned to their corresponding species, revealing a low rate of doublet formation ($n = 4{,}631$ human cells, 312 mouse cells, 36 mixed). **i**, Mean CellTags per cell for human and mouse cells in the species-mixing experiment. CellTag transcripts were detected in 70% of cells ($n = 3{,}493/4{,}979$ cells). Of the tagged population, each cell expressed 5 CellTags on average: 3.800 ± 0.002 in human cells, and 5.90 ± 0.02 in mouse cells (mean ± s.e.m.). **j**, For each cell, CellTag signatures were extracted and Jaccard similarity analysis was performed to assess the frequency of CellTag signature overlap between the two species. To establish a false-positive baseline, we initially compared CellTag overlap between mouse and human populations, as these cells are not related. From the analysis of 4,943 cells, we identified 200 instances of mouse–human cell pairings out of a possible 1.5×10^7 pairs sharing the same individual CellTags. This demonstrates that reliance on only one CellTag per cell does not uniquely label cells with high confidence. Excluding cells represented by only one CellTag removes this noise, resulting in no detection of cross-species CellTag signatures (Jaccard similarity index <0.7). This highlights the importance of combinatorial labelling, and the efficacy of our approach to uniquely label unrelated cells.

© 2018 Springer Nature Limited. All rights reserved.

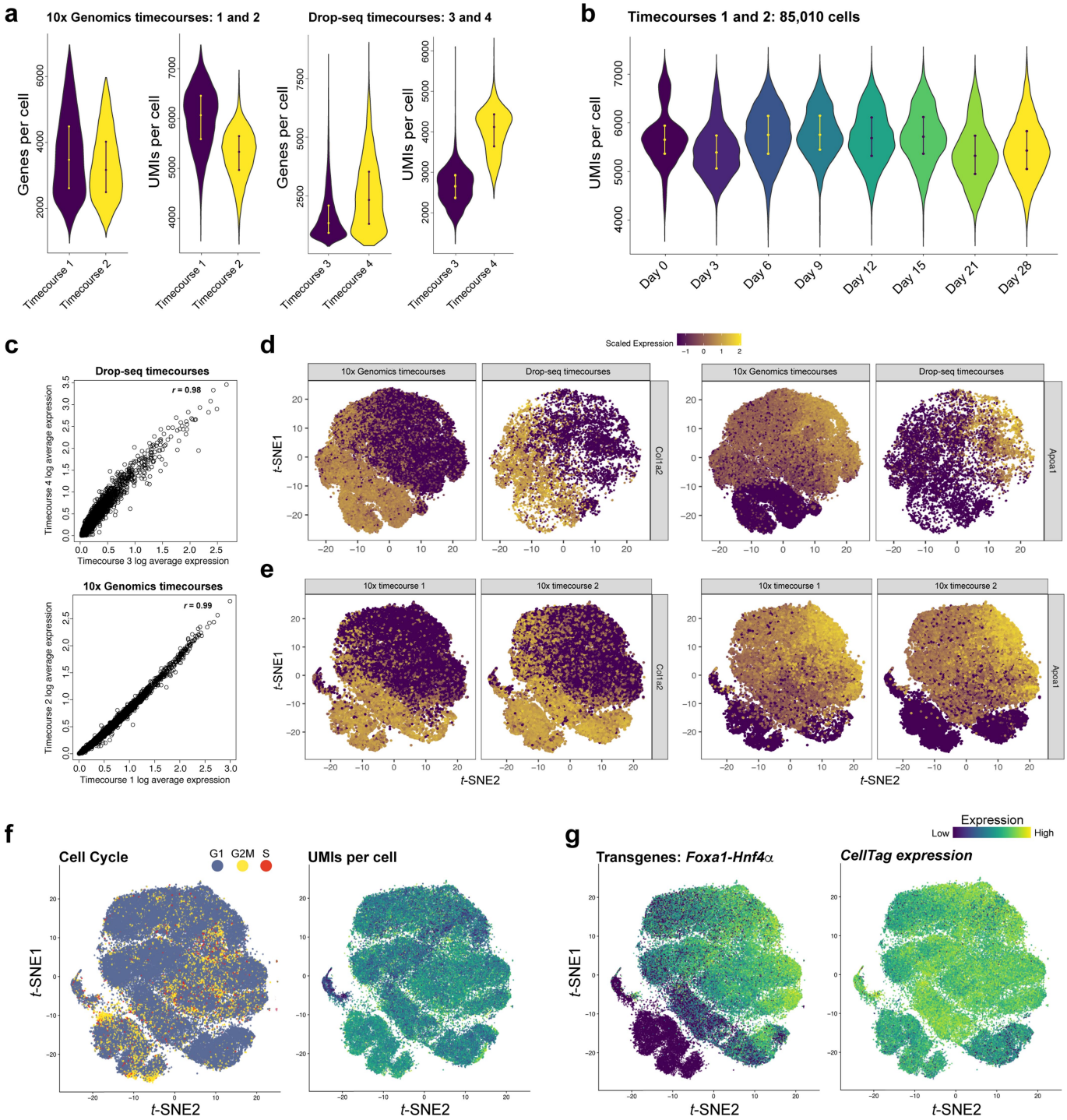


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

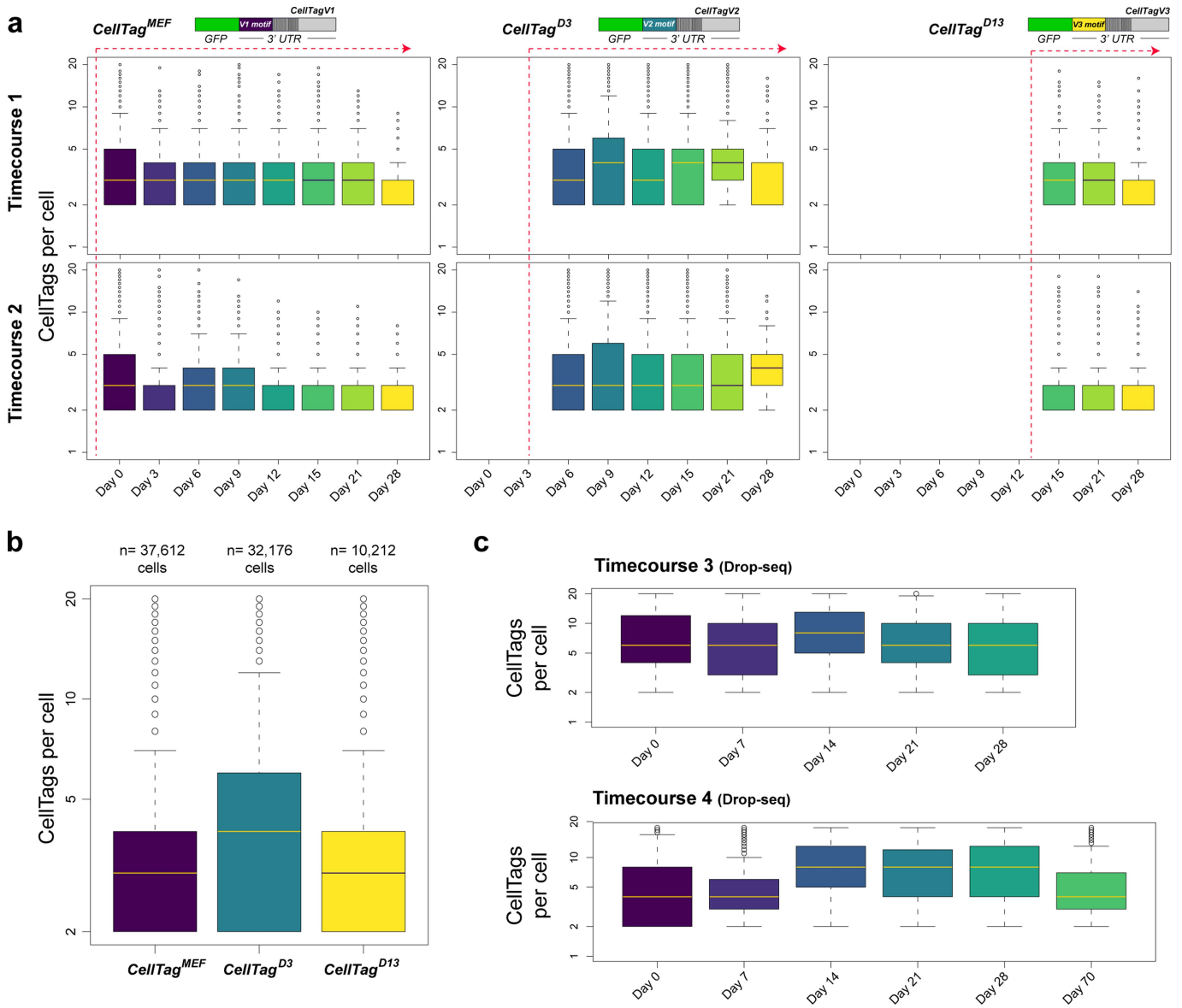
Extended Data Fig. 2 | CellTagging does not perturb cell physiology or reprogramming efficiency. To assess the potential effect of CellTagging on cell physiology we performed scRNA-seq on CellTag-labelled cells and unlabelled control cells 72 h after tagging. **a**, Left, fluorescent image of CellTag-labelled, GFP-expressing, pre-B cell line, HAFTL-1. Right, 10x Genomics-based scRNA-seq of CellTag-labelled ($n = 3{,}943$ cells) and non-tagged control cells ($n = 2{,}067$ cells). Cells were clustered using Seurat, resulting in a *t*-SNE plot with 6 clusters of transcriptionally distinct cells. CellTag-labelled and control cells were evenly distributed across these populations. **b**, The CellTag-labelled B-cell population expresses a mean of 3.50 ± 0.02 CellTags per cell. **c**, We detect no observable differences in numbers of genes or UMIs per cell in either population. **d**, Average gene expression values between CellTag-labelled and control cells are highly correlated ($r = 0.999$, Pearson's correlation), demonstrating that our labelling approach does not induce significant changes in gene expression. These experiments were performed independently twice with similar results. **e**, To assess the potential effect of CellTagging on reprogramming outcome, we induced lineage conversion (MEF to iEP) of CellTagged cells in parallel with unbarcoded control cells, followed by three weeks of culture and processing on the Drop-seq platform ($n = 773$ cells passing quality control). A mean of 3.30 ± 0.09 CellTags per cell are expressed in a labelled reprogrammed cell population. **f**, There are no observable differences in numbers of genes or UMIs per cell in either the labelled or unlabelled populations. **g**, Average gene expression values between CellTagged and control cells are highly correlated ($r = 0.98$, Pearson's correlation), again demonstrating that our labelling approach does not induce significant changes in gene expression. **h**, Seurat clustering of cells, in which cells in fibroblast (*Col1a2*-high), transition, and fully reprogrammed (*Apoa1*-high) states can be identified. Right, barcoded and control cells are distributed fairly evenly across these reprogramming stages. Some variation is expected between these independent biological replicates. These experiments were performed independently twice with similar results.

© 2018 Springer Nature Limited. All rights reserved.



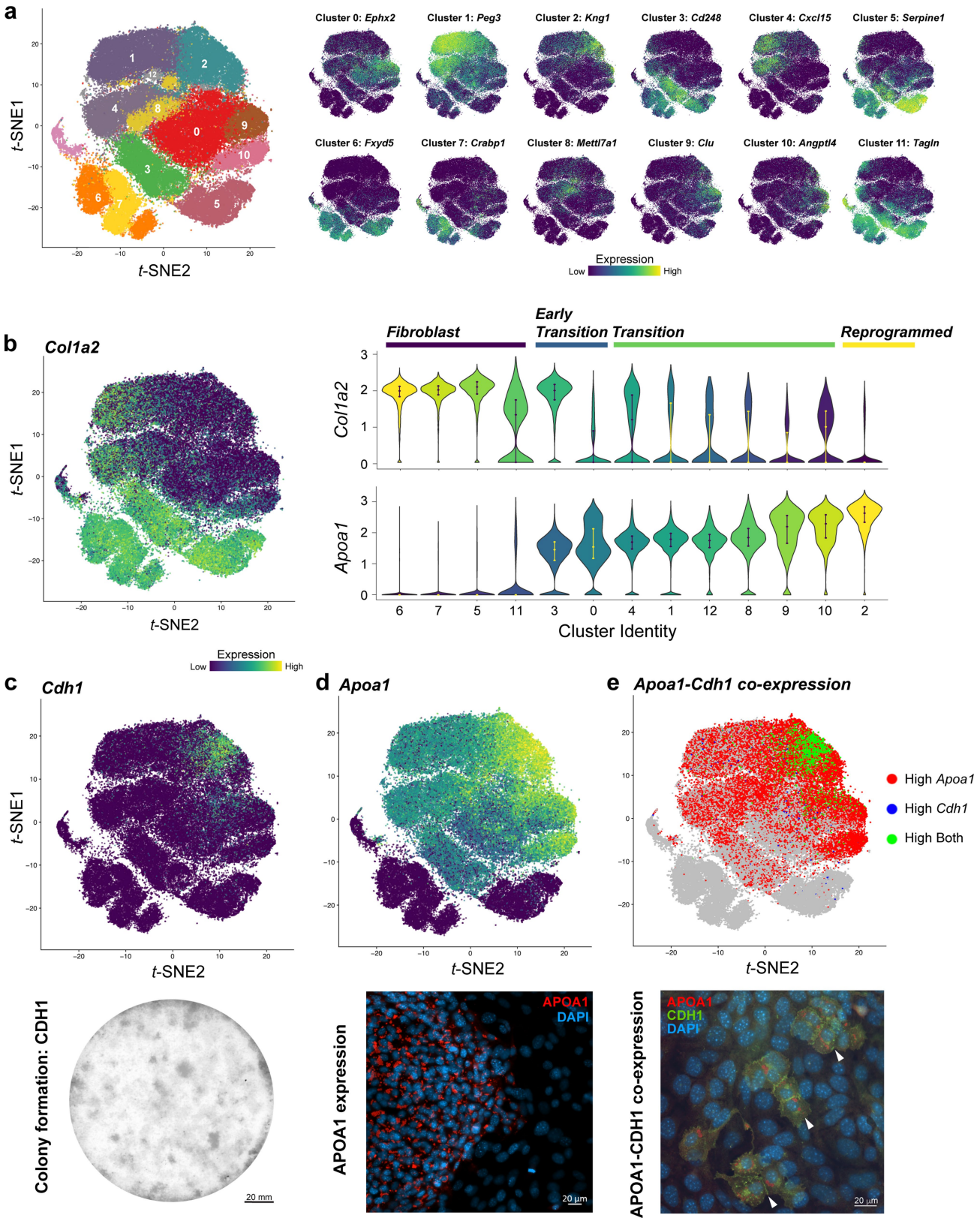
Extended Data Fig. 3 | scRNA-seq metrics and quality control of cell clustering. a, Numbers of genes and UMIs per cell for 10x Genomics-based (time course 1, $n = 30{,}733$ cells and time course 2: $n = 54{,}277$ cells) and Drop-seq-based (time course 3, $n = 5{,}932$ cells and time course 4: $n = 5{,}414$ cells) reprogramming time courses. In these cross-platform comparisons, we apply more stringent filtering of Drop-seq data to include only those cells with 1,000 or more UMIs. For Drop-seq experiments, with a cell capture rate of 5%, 2×10^5 MEFs were initially seeded for reprogramming. For 10x Genomics experiments, with a cell encapsulation rate of up to 60%, 5×10^4 MEFs were initially seeded for reprogramming. **b**, Mean numbers of UMIs per cell at each captured time point during reprogramming (5,570.0 ± 2.2), in two independent biological replicates (10x Genomics, time courses 1 and 2): cells were captured at days 3, 6, 9, 12, 15, 21 and 28, along with the initial MEF population (day 0). **c**, Average gene expression values of 10x Genomics and Drop-seq replicates are highly correlated at day 0, demonstrating technical consistency ($r = 0.99$, and $r = 0.98$, respectively, Pearson's correlation). **d**, Alignment of independent 10x Genomics replicates (time courses 1 and 2) with Drop-seq replicates (time courses 3 and 4) using canonical correlation analysis[19]. Left, expression of MEF marker *Col1a2*. Right, iEP marker *Apoa1*. Overlay of data from these two sources demonstrates a high level of technical and biological consistency between the two technologies. **e**, Alignment of 10x Genomics replicates (time course 1 and 2) using canonical correlation analysis. Expression of *Col1a2* (left), *Apoa1* (right). Integration of these two replicates demonstrates a high level of technical and biological consistency. **f**, Projections of cell cycle phase and UMIs per cell onto *t*-SNE alignment of time courses 1 and 2 shows that clustering is independent of these factors. **g**, Reprogramming factor expression (using detection of bicistronic Hnf4a-T2A-Foxa1 transgene expression) and CellTag expression across time courses 1 and 2.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | CellTag expression metrics. a, Mean counts of CellTags expressed per cell, following whitelisting and filtering for time course 1 ($n = 19{,}581$ cells passing filtering) and 2 ($n = 38{,}943$ cells passing filtering), broken down by time point and CellTag version. Red dashed lines denote time of CellTag transduction. **b**, Mean number of CellTags expressed per cell, post-whitelisting and filtering, for each round of barcoding across time courses 1 and 2. CellTag^{MEF}: 3.40 ± 0.01 CellTags per cell, $n = 37{,}612$ cells; CellTag^{D3}: 4.50 ± 0.02 CellTags per cell, $n = 32{,}176$ cells; CellTag^{D13}: 3.20 ± 0.02 CellTags per cell, $n = 10{,}212$ cells. Sixty-five per cent of sequenced cells pass the ≥ 2 CellTag expression threshold to support tracking. **c**, Mean CellTags per cell following whitelisting and filtering for both Drop-seq time courses, broken down by time point. All cells with 200 or more genes were included in this analysis (time course 1: $n = 10{,}038$ cells; time course 2: $n = 9{,}839$ cells). CellTags were introduced only in MEFs, before reprogramming in these experiments. In Drop-seq time courses, we detected a mean of 7.80 ± 0.07 CellTags per cell, across 61% of cells (12,086/19,877 cells) passing the tracking threshold.

© 2018 Springer Nature Limited. All rights reserved.

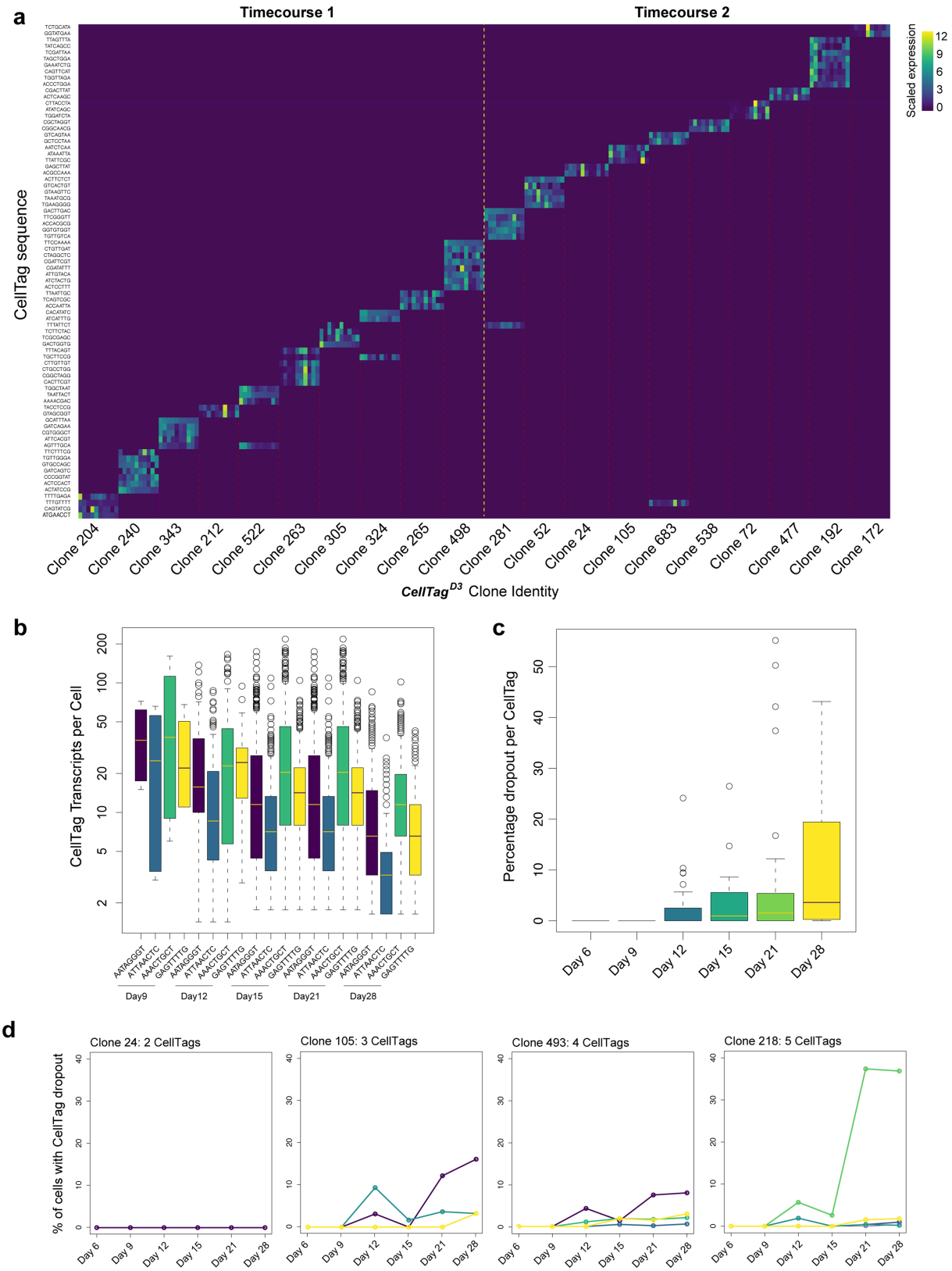


Extended Data Fig. 5 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 5 | Assignment of cluster identities based on mRNA and protein expression. a, Top enriched gene expression associated with each cluster, projected onto the reprogramming *t*-SNE plot ($n = 85{,}010$ cells). **b**, Left, expression of *Col1a2*, projected onto the *t*-SNE plot. Top right, violin plot of *Col1a2* expression levels in each cluster. Bottom right, violin plot of *Apoa1* expression levels in each cluster, ordered by gain of expression over the course of reprogramming. Clusters are classified as one of four reprogramming stages: fibroblast, clusters 5, 6, 7, 11; early transition, clusters 0, 3; transition, clusters, 1, 4, 8, 9,10, 12; and reprogrammed, cluster 2. *Apoa1* is not expressed in the fibroblast clusters. **c**, Top, expression of the iEP marker[3,16] *Cdh1* (E-cadherin), projected onto the *t*-SNE plot, highlighting the location of fully reprogrammed cells. Bottom, staining of CDH1 protein in iEP colonies emerging following three weeks of reprogramming (control shown is from Fig. 4d). Scale bar, 20 mm. **d**, Top, expression of the novel iEP marker, apolipoprotein A1, *Apoa1*, projected onto the *t*-SNE plot. Bottom, immunofluorescence of APOA1 protein in an iEP colony, following three weeks of reprogramming. APOA1 (red) is localized to vesicles. This is a representative image selected from five independent biological replicates. Scale bar, 20 μm. **e**, Top, co-expression of *Apoa1* and *Cdh1* at the transcript level within the same individual cells in the fully reprogrammed cluster confirms *Apoa1* as a marker of iEP emergence. Bottom, immunofluorescence of APOA1 and CDH1 protein in iEPs. White arrows mark emerging iEP colonies co-expressing both proteins. APOA1 expression (red) is found localized to vesicles of CDH1-positive cells (green), where the most intense CDH1 staining is observed at cell–cell junctions. This is a representative image selected from three independent biological replicates. Scale bar, 20 μm.

© 2018 Springer Nature Limited. All rights reserved.

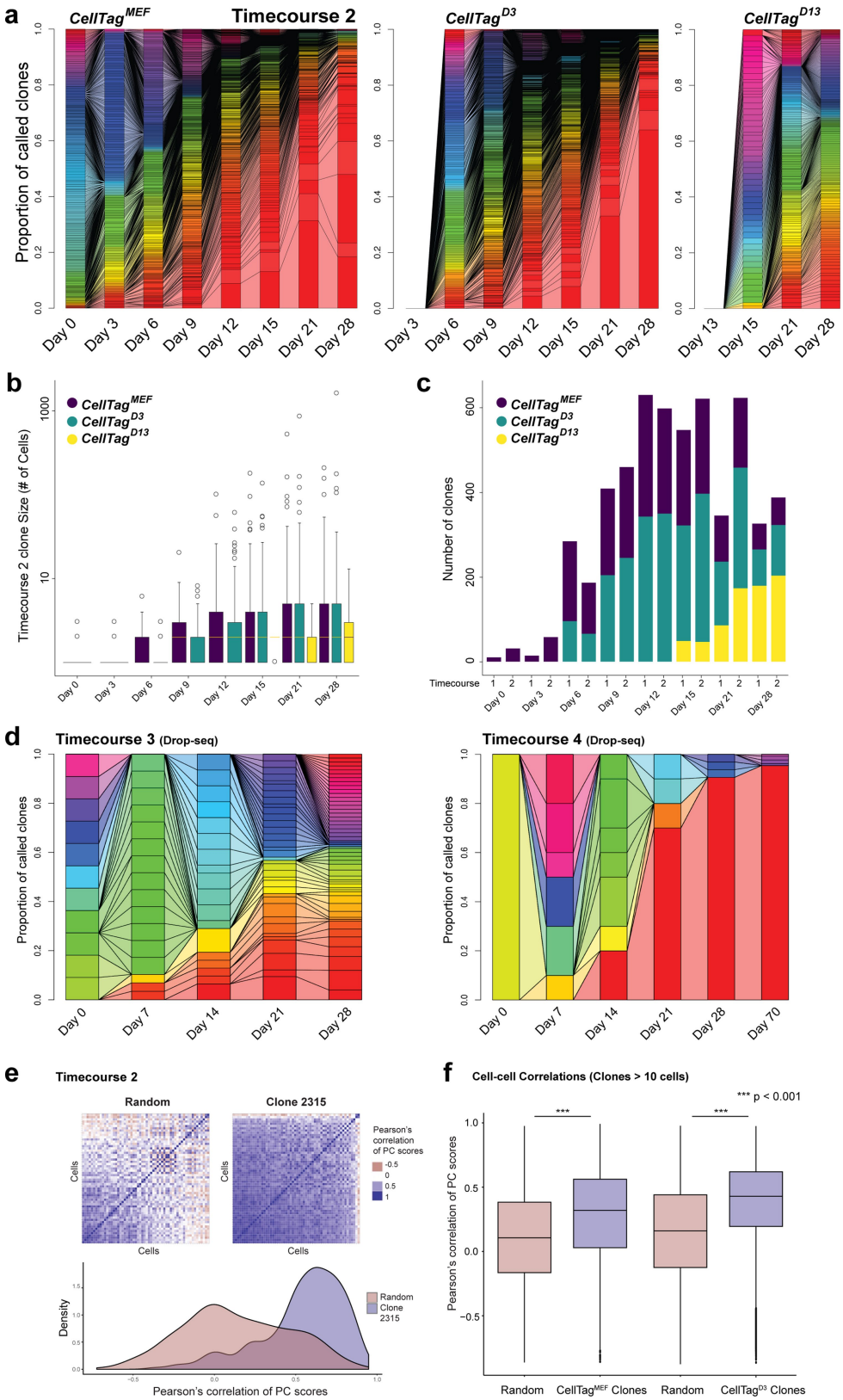


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 6 | Combinatorial CellTag labelling to identify clonally related cells. a, Heat map showing scaled expression of individual CellTags in 20 major clones from cells labelled with $CellTag^{D3}$ ($n = 10$ representative cells per clone, time courses 1 and 2). The dashed yellow line marks separation between the two time courses. Dashed red lines mark separation between independent clones. Although some CellTags are shared between these independent biological replicates, the combined CellTag signatures are unique. **b**, Expression levels of individual CellTags per cell over three weeks in a representative clone labelled by four unique CellTags. Expression diminishes over time, but is not completely silenced. **c**, To assess CellTag silencing, we selected 10 major clones ($n = 6{,}728$ cells), defining the intact CellTag signature for each clone at reprogramming day 6. We then assessed loss, or 'dropout' of CellTags from each signature over the time course to day 28. By week 4, expression of an individual CellTag is lost in 1 out of 10 cells—that is, expected CellTag expression was not detected in $11 \pm 2\%$ of cells. Conversely, CellTag expression is retained in almost 90% of cells by day 28. Later rounds of CellTag labelling ($CellTag^{D13}$) are less prone to this effect, with CellTags dropping out in only $3.0 \pm 1.5\%$ of cells. **d**, We mapped CellTag expression across four representative clones, in which expression of each CellTag is plotted over time. The *y* axis denotes the percentage of cells within each clone in which expression of specific CellTags has dropped out. Typically, only one CellTag exhibits dropout, and expression of the other CellTags is maintained. We do not observe complete silencing, that is, loss of expected CellTag expression in 100% of cells. This demonstrates the advantage of our CellTag combinatorial indexing method to reliably label cells and track them over an extended period of time. For example, reliance on the expression of a single, longer barcode would not be effective following integration into a region that later becomes silenced.

© 2018 Springer Nature Limited. All rights reserved.

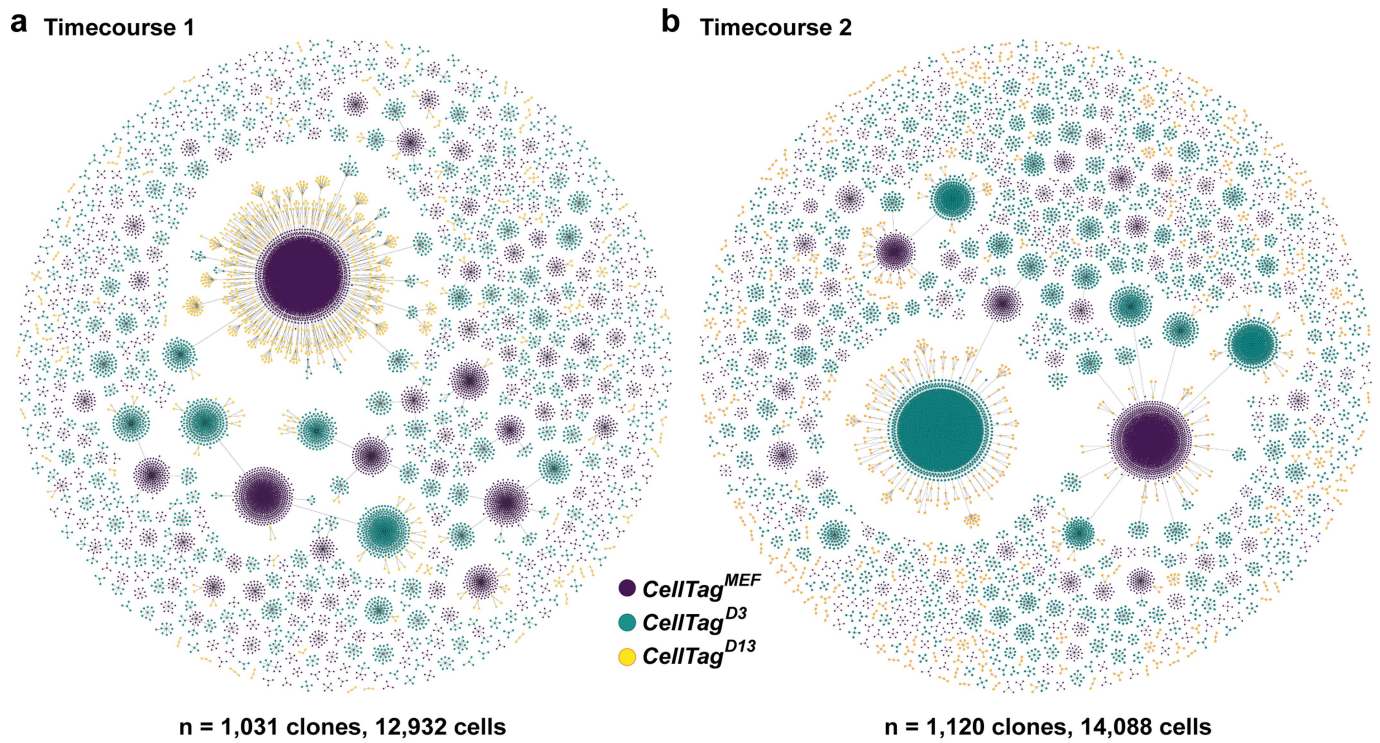


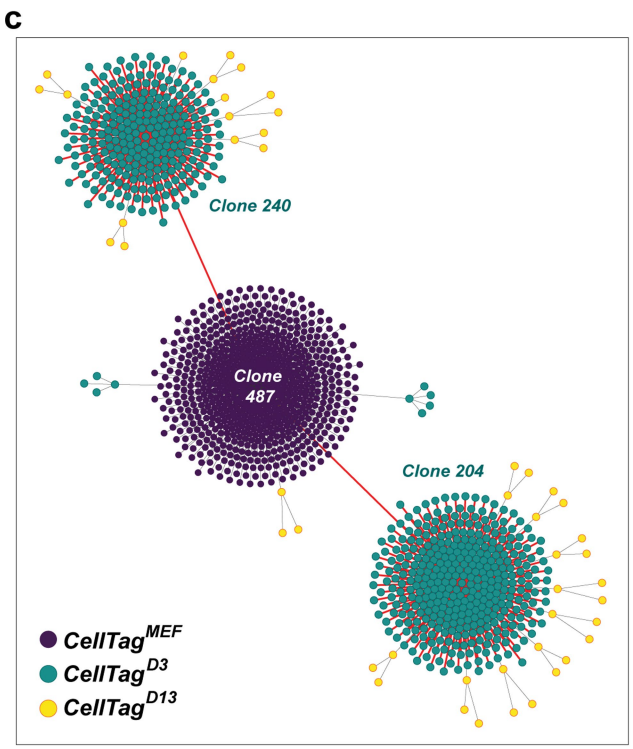
Extended Data Fig. 7 | See next page for caption.

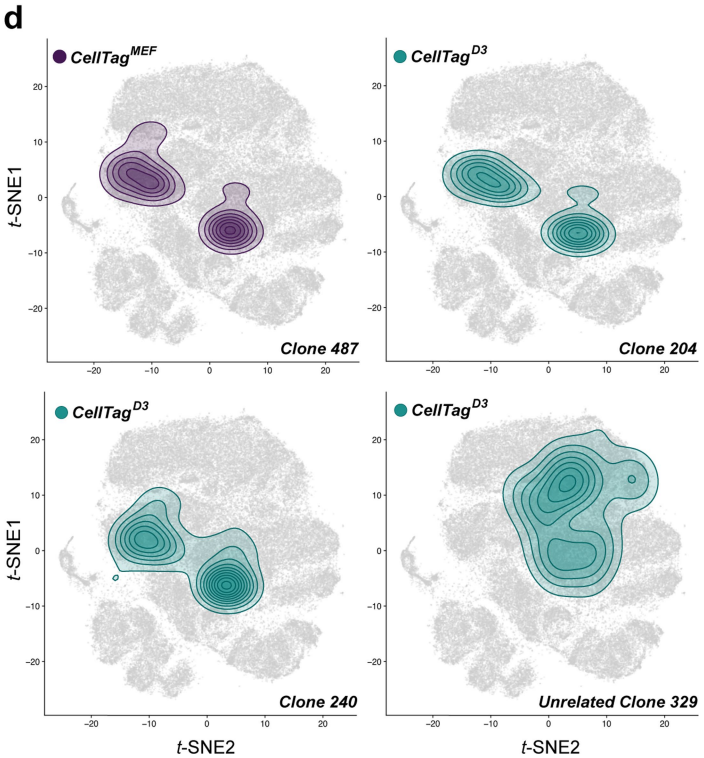
© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 7 | Visualizing growth of clones and gene expression correlation within clones. a, Connected bar plots showing individual clones as a proportion of all clones at each reprogramming time point for time course 2, for each round of CellTagging ($n = 14{,}088$ cells across 1,120 clones). Connected bars denote clonal expansion and growth over time. **b**, Average number of cells per clone, per time point, for each round of CellTag labelling (time course 2, $n = 1{,}120$ clones). **c**, Number of clones detected at each time point, for each round of CellTagging over reprogramming time courses 1 ($n = 1{,}031$ clones) and 2 ($n = 1{,}120$ clones). The number of clones detected gradually increases over time as the probability of capture increases with clonal growth. The number of clones then begins to decrease as the growth of some individual clones out-competes other clones, which are lost from the population over time. **d**, Connected bar plots showing individual clones as a proportion of all clones called at each reprogramming time point for Drop-seq replicate 1 ($n = 103$ clones) and Drop-seq replicate 2 ($n = 37$ clones). In replicate 2, a single clone progressively dominates the culture over 10 weeks of growth. In our viral integration analyses (Supplementary Table 5), we detect three viral integration sites in the cells of this clone. We did not detect any differential expression of genes proximal to these integration sites. Similarly, analysis of gene expression enrichment in 12 dominant clones across two biological replicates does not reveal any common signature of these clones to explain their rapid expansion (data not shown). This suggests that the clonal growth we observe is a normal part of the iEP reprogramming process, in which the cells enter a progenitor-like state. Even so, these analyses do not exclude the acquisition of genetic and epigenetic changes endowing these expanding clones with increased fitness. **e**, Correlation of principal component (PC) scores in clonally related cells (clone 2315, $n = 58$ cells) relative to a random sampling of cells. Correlation between PC scores was used as a proxy for transcriptional similarity between cells. Clonally related cells were much more closely correlated, relative to randomly selected cells. **f**, Quantification of correlation analysis for all time course 2 clones consisting of 10 cells or more, for $CellTag^{MEF}$ ($n = 78$ clones, 3,963 cells) and $CellTag^{D3}$-labelled clones ($n = 109$ clones, 6,265 cells). Mean correlation scores for clonally related cells are significantly higher than random cell groupings ($P < 0.001$, *t*-test, one-sided). We tagged cells both before and after the 72-h reprogramming window, expecting substantial heterogeneity to be introduced by serial viral transduction. On the contrary, there is only a slight but insignificant increase in PC score correlation between $CellTag^{MEF}$ and $CellTag^{D3}$-labelled, clonally related cells.

© 2018 Springer Nature Limited. All rights reserved.

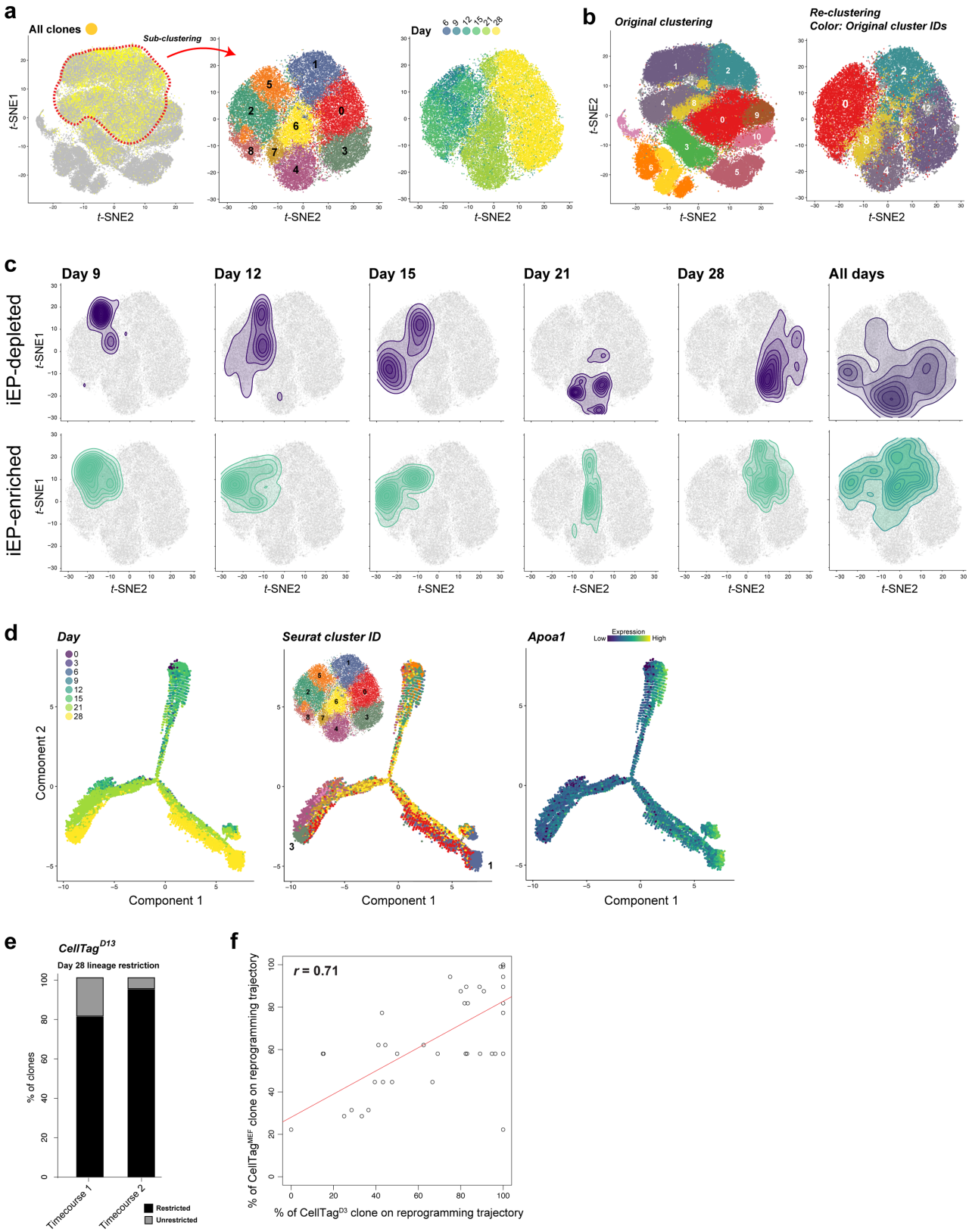






Extended Data Fig. 8 | Reconstruction and visualization of lineages via force-directed graph drawing. a, **b**, Force-directed graph of all clonally related cells and lineages reconstructed from time course 1 (1,031 clones, 12,932 cells) (**a**) and time course 2 (1,120 clones, 14,088 cells) (**b**). All lineages and clone distributions can be interactively explored via our companion website, CellTag Viz (http://www.celltag.org/). **c**, In this tree, we follow CellTag^{MEF} clone 487 from time course 1 and its descendants. Each node represents an individual cell, and edges represent clonal relationships between cells. Purple, CellTag^{MEF} clones; blue, CellTag^{D3} clones; yellow, CellTag^{D13} clones. In the lineage highlighted in red, we follow the CellTag^{MEF} clone ($n = 678$ cells), branching into two CellTag^{D3} lineages (clone 204 ($n = 363$ cells) and clone 240 ($n = 260$ cells)). **d**, Contour plots, representing cell density of each clone, projected onto the *t*-SNE plot, for the lineage shown in **c**. Top left, cells belonging to clone 487 (CellTag^{MEF}). Clones 204 and 240 (CellTag^{D3}) descend from this first clone, exhibiting a high degree of overlap within 2D space, on the *t*-SNE plot. An unrelated CellTag^{D3} clone, 329 ($n = 38$ cells), does not overlap with this lineage, demonstrating the high degree of similarity between cells belonging to the same lineage.

© 2018 Springer Nature Limited. All rights reserved.

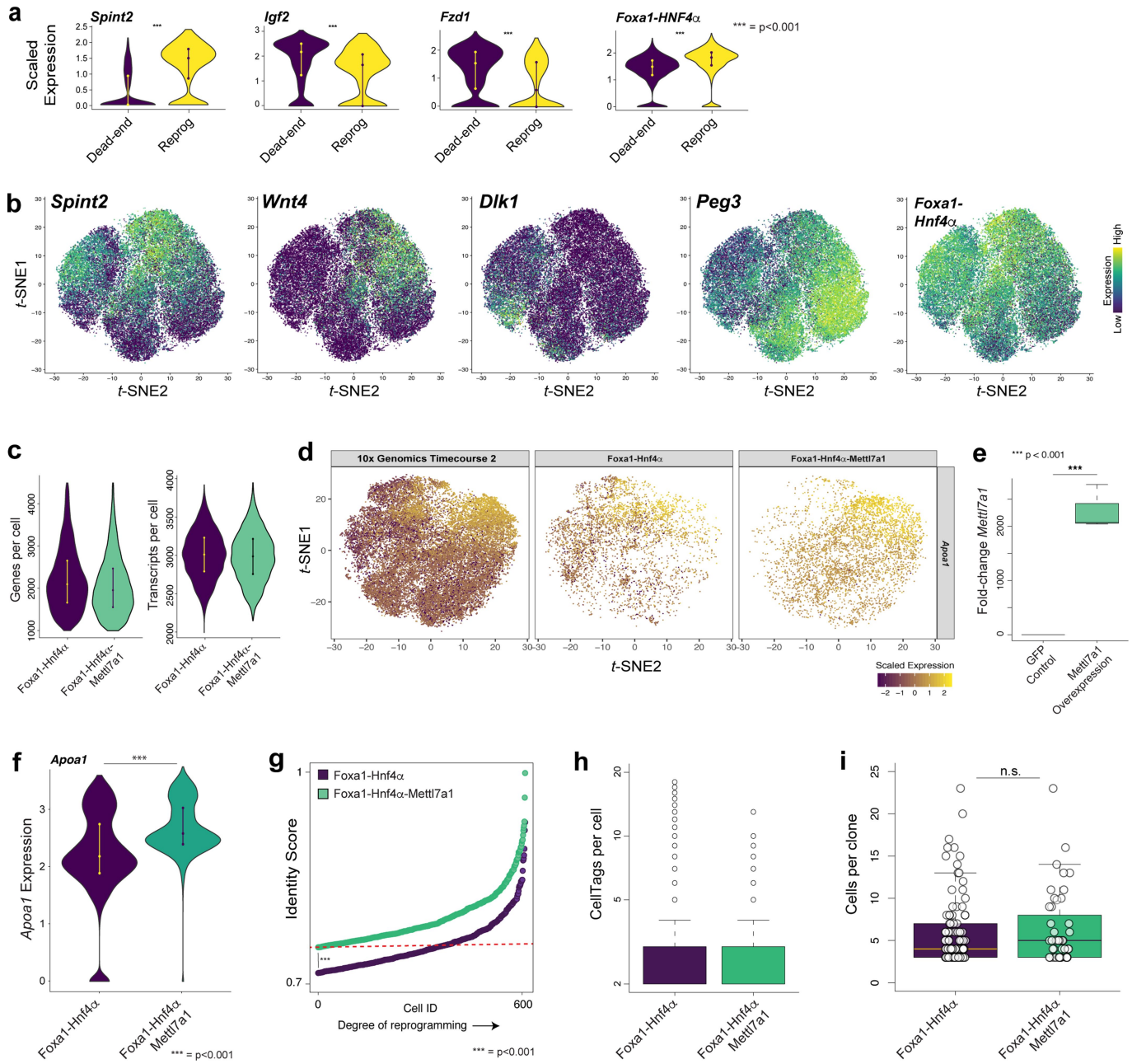


Extended Data Fig. 9 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | Mapping reprogramming trajectories and timing of cell fate decisions. a, Projection of all clones (yellow, $n = 2{,}151$ clones, 27,020 cells) across reprogramming time courses 1 and 2 ($n = 85{,}010$ cells). A subset of clusters with the highest density of detected clones, outlined in red (clusters 0, 1, 2, 4, 8, and 12), were extracted from this larger dataset and re-clustered to generate a higher-resolution *t*-SNE plot, focusing on reprogramming days 6 to 28 ($n = 48{,}515$ cells). **b**, Left, original cluster identities of all cells ($n = 85{,}010$ cells). Right, subset of 48,515 cells, coloured by original cluster identity. **c**, Contour plots of iEP-depleted clone distribution (top panels, ($n = 7$ clones, 1,037 cells)) and iEP-enriched clone distribution (bottom panels, ($n = 7$ clones, 2,270 cells)) broken down by reprogramming day, and across days 9–28 (far right). These specific clones were selected from the larger iEP-depleted and iEP-enriched groups, as they included cells distributed across all time points, enabling their trajectories to be defined. In these distributions, clusters 8, 4 and 3 are iEP-depleted, thus representing the dead-end trajectory. Conversely, clusters 2, 6 and 1 are iEP-enriched, representing the reprogramming trajectory. These trajectories divide cluster 0 into two halves, but re-clustering does not increase resolution (data not shown). Deeper sequencing of a larger number of cells may provide further insights into this cluster in future studies. **d**, Monocle2 psuedotemporal ordering of cells in the subset of cells ($n = 48{,}515$ cells), coloured by day of reprogramming (left panel), Seurat cluster ID (middle panel) and *Apoa1* expression (right panel). Monocle2 uses dimension reduction to represent each single cell in 2D space and effectively 'connects the dots' to construct a reprogramming trajectory. In this analysis, we performed semi-supervised ordering using *Col1a2* (marking fibroblast identity) expression as a start point and *Apoa1* expression (marking iEP identity) as an endpoint. The branched trajectory generated by Monocle2 is in general agreement with our clonal analyses. **e**, Restriction of $CellTag^{D13}$ clones (time course 1, $n = 79$ clones, 240 cells; time course 2, $n = 30$ clones, 148 cells) to either the reprogrammed cluster (cluster 1), or the dead-end cluster (cluster 3) at day 28. Of the clones from these two biological replicates, $88 \pm 8\%$ exhibit adherence to one of these trajectories by day 13 of reprogramming. **f**, We identified lineages in which multiple $CellTag^{D3}$-labelled clones share a common $CellTag^{D0}$-labelled ancestor. The proportion of each clone on the reprograming trajectory (defined as occupancy of clusters 2, 6 and 1 on the t-SNE plot of the subset of clusters), and proportion of each clone on the dead-end trajectory (defined as occupancy of clusters 8, 4 and 3) was calculated. We then plotted the proportion of each $CellTag^{MEF}$-labelled clone on the reprogramming trajectory against that of its $CellTag^{D3}$-labelled descendants ($r = 0.71$, Pearson's correlation, $n = 13$ lineages, 57 clones, 6,035 cells).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | *Mettl7a1* expression is upregulated on the reprogramming trajectory, and promotes iEP generation. a, Violin plots of significantly different gene expression between reprogramming and dead-end trajectories ($n = 2{,}074$ cells). **b**, Projection of gene expression onto the *t*-SNE plot ($n = 48{,}515$ cells). *Wnt4* and *Spint2* expression is significantly upregulated along the reprogramming trajectory ($P < 0.001$, permutation test, one-sided, $n = 1{,}037$ cells). *Dlk1* and *Peg3* expression is significantly upregulated along the dead-end trajectory ($P < 0.001$, permutation test, one-sided, $n = 1{,}037$ cells). Expression of the Foxa1-Hnf4a transgene is significantly downregulated along the dead-end trajectory ($P < 0.001$, permutation test, one-sided, $n = 1{,}037$ cells). **c**, Mean numbers of genes and transcripts per cell following 10x Genomics-based scRNA-seq analysis: Foxa1-Hnf4a reprogrammed cells ($n = 6{,}559$ cells) and Foxa1-Hnf4a-Mettl7a1 reprogrammed cells ($n = 10{,}161$ cells), collected 14 days after initiation of reprogramming. For subsequent analyses, the Foxa1-Hnf4a-Mettl7a1 experimental group was randomly downsampled for direct comparison to the Foxa1-Hnf4a experimental group ($n = 6{,}559$ cells for both groups). **d**, The Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1 scRNA-seq datasets were merged with cells from time course 2, using canonical correlation analysis[19], to help place these two experimental groups on the previously defined trajectories. Expression levels of *Apoa1* are projected onto this *t*-SNE plot. **e**, Confirmation of *Mettl7a1* expression by qRT–PCR, following transduction of cells with Foxa1-Hnf4a-GFP versus Foxa1-Hnf4a-Mettl7a1 retroviruses ($^{**}P = 5.3 \times 10^{-3}$, *t*-test, one-sided). **f**, Violin plot of mean *Apoa1* expression in cells reprogrammed with Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1. Addition of Mettl7a1 to the reprogramming cocktail results in a significant increase in *Apoa1* expression, supporting observations that this factor increases the yield of fully reprogrammed cells ($P < 0.001$, permutation test, one-sided). **g**, Plot of identity scores of Foxa1-Hnf4a (purple) and Foxa1-Hnf4a-Mettl7a1 (green) reprogrammed cells. Cells are ordered according to an increase in iEP identity. Red dashed line indicates a cut-off of 0.75; above this score cells are considered as iEPs. Threefold-more Foxa1-Hnf4a-Mettl7a1 cells classify as iEPs, relative to Foxa1-Hnf4a cells, represented as a significant increase in iEP score ($P < 0.001$, permutation test, one-sided). **h**, Box plot of mean CellTag expression between Foxa1-Hnf4a (3 ± 0.05 CellTags per cell) and Foxa1-Hnf4a-Mettl7a1 (2.5 ± 0.04 CellTags per cell) experimental groups. The box plots show the median, first and third quantile, and error bar with outliers. **i**, Box plot of cells per clone for Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1 experimental groups, following data processing via our CellTag demultiplexing and clone calling pipeline. Clone size does not significantly differ between these two groups: Foxa1-Hnf4a, 6.0 ± 0.4 cells per clone ($n = 99$ clones, 595 cells); Foxa1-Hnf4a-Mettl7a1: 6.30 ± 0.65 cells per clone ($n = 43$ clones, 277 cells), demonstrating that the addition of Mettl7a1 enhances iEP yield by increasing the number of unique reprogramming events. For comparison, average clone size at ~day 14 for time course replicates 1 and 2 is ~8 cells per clone.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0766-y

An entanglement-based wavelength-multiplexed quantum communication network

Sören Wengerowsky[1,2]*, Siddarth Koduru Joshi[1,2,4], Fabian Steinlechner[1,2,5,6], Hannes Hübel[3] & Rupert Ursin[1,2]*

Quantum key distribution[1] has reached the level of maturity required for deployment in real-world scenarios[2–6]. It has previously been shown to operate alongside classical communication in the same telecommunication fibre[7–9] and over long distances in fibre[10,11] and in free-space links[12–15]. Despite these advances, the practical applicability of quantum key distribution is curtailed by the fact that most implementations and protocols are limited to two communicating parties. Quantum networks scale the advantages of quantum key distribution protocols to more than two distant users. Here we present a fully connected quantum network architecture in which a single entangled photon source distributes quantum states to many users while minimizing the resources required for each. Further, it does so without sacrificing security or functionality relative to two-party communication schemes. We demonstrate the feasibility of our approach using a single source of bipartite polarization entanglement, which is multiplexed into 12 wavelength channels. Six states are then distributed between four users in a fully connected graph using only one fibre and one polarization analysis module per user. Because no adaptations of the entanglement source are required to add users, the network can readily be scaled to a large number of users, without requiring trust in the provider of the source. Unlike previous attempts at multi-user networks, which have been based on active optical switches and therefore limited to some duty cycle, our implementation is fully passive and thus has the potential for unprecedented quantum communication speeds.

The quantum key distribution (QKD) networks that have been demonstrated so far can be grouped into five types of configuration. First, quantum repeater networks use quantum memories and entanglement swapping to extend and route quantum states and to form arbitrary network topologies. Although quantum repeaters are very likely to feature prominently in future quantum networks, technological advancement in quantum memories is needed for quantum repeater networks to be considered practical. However, quantum repeaters can also be used to improve the performance of other types of quantum network.

The second type of configuration uses high-dimensional or multipartite entanglement to share entanglement resources between several users[16]. This way, different users share different subspaces of the Hilbert space to generate their keys. However, adding or removing users requires changes in the dimensionality of the system, which makes complex alterations of the source necessary.

The third type of configuration is trusted node networks. They amount to a mesh of point-to-point links, each requiring a complete two-party communication set-up. Although trusted nodes have been used to extend bipartite quantum communication schemes to larger multi-user networks[2–5], they also relinquish the strong security offered by quantum cryptography. Furthermore, this approach creates a substantial resource overhead because it duplicates sender and receiver hardware.

The fourth type of configuration realizes a point-to-multipoint network consisting of two (or more) sets of users, in which a member of the first set can communicate with any member of the second set but not with members of the same set. This type of configuration allows multiple users to share receivers or sources and has been realized in configurations with passive beam splitters[7,17,18], active optical switches that establish a temporary quantum channel between two particular users at a time[6,19–21], and frequency multiplexing[21–24].

The final type of configuration—the most versatile and robust architecture—is a fully connected network architecture connecting every user to every other user simultaneously. A reconfigurable point-to-multipoint network, in which a user can request to be connected to any other user one at a time, has been used to achieve some of the benefits of a fully connected network[25].

Here we present a fully connected network architecture and its realization in the telecommunications band without any requirement for active switching. The transition to all-passive optical networks offers a substantial boost in terms of reliability and miniaturization. Further, it does not limit the distribution rate as per the duty cycle of the switching device. We connected four simultaneously active users to a polarization-entangled photon source via a single optical fibre each. Using the frequency correlations of the photons via wavelength-division multiplexing (WDM), we distributed bipartite entanglement between all pairs of users. This allows all pairs of users to generate their own private key using only a single source.

The complete network architecture can be better understood if divided into layers of abstraction (Fig. 1). The bottom ('physical') layer contains all of the tangible components (physical connections) and forms the physical topology of the network. Each of the four users (Alice, Bob, Chloe and Dave) receives a combination of three wavelength channels via a single-mode fibre. Thus, the source distributes six bipartite entangled photon states to the four users. The middle ('quantum correlation') layer represents the six entangled states (which each corresponds to a different secure key) that link the four users. The top ('communication') layer represents secure classical communication between users and the logical topology of the network.

To create a fully connected graph in the quantum correlation and communication layers with N users, we need a minimum of $N(N-1)/2$ links. Each of the N users is equipped with a single detection module, exactly the same as in standard two-party quantum communication schemes. The service provider multiplexes $N-1$ channels into each single-mode fibre. Thus, using $N(N-1)$ channels, $N(N-1)/2$ entangled photon pairs (and hence secure keys) can be shared by any pair of users. As the number of users increases, the physical topology of the resulting network remains elementary and grows linearly because all channels needed by each user are multiplexed into the same single-mode fibre and detection system. However, the logical topology (that is, the structure and number of quantum correlations and communication links) becomes increasingly complex and grows quadratically. This scalability allows us to easily create large complex networks

[1]Institute for Quantum Optics and Quantum Information—Vienna, Austrian Academy of Sciences, Vienna, Austria. [2]Vienna Center for Quantum Science and Technology, Faculty of Physics, University of Vienna, Vienna, Austria. [3]Austrian Institute of Technology, Vienna, Austria. [4]Present address: Quantum Engineering Technology Labs, H. H. Wills Physics Laboratory, Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK. [5]Present address: Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany. [6]Present address: Friedrich Schiller Universitz Jena, Abbe Center for Photonics, Jena, Germany. *e-mail: Soeren.Wengerowsky@oeaw.ac.at; Rupert.Ursin@oeaw.ac.at

© 2018 Springer Nature Limited. All rights reserved.

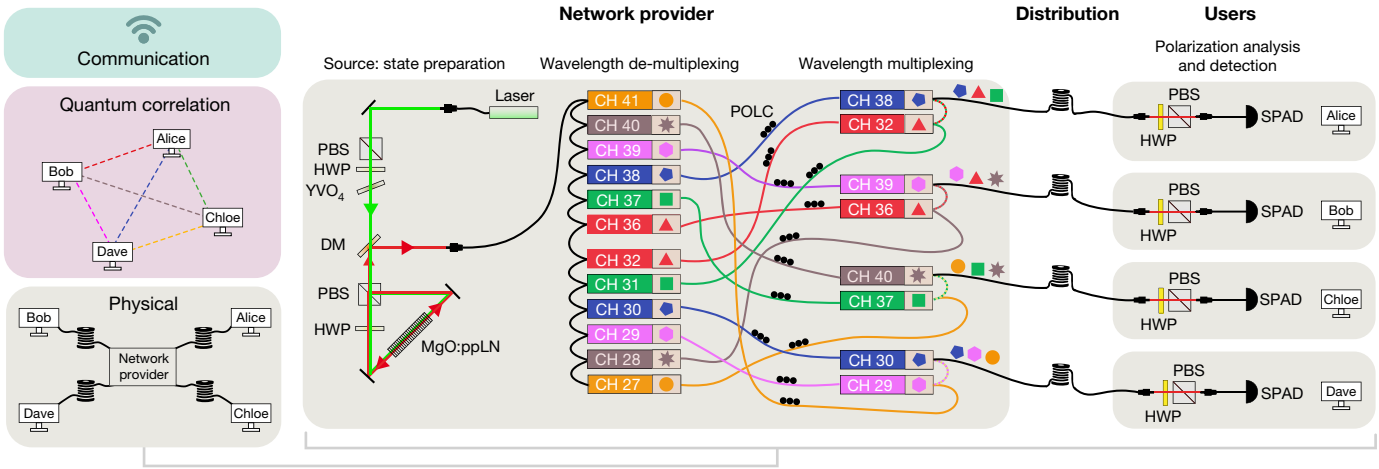


Fig. 1 | Network architecture and experimental set-up. On the left we illustrate the network architecture using three layers of abstraction. The bottom layer represents the physical topology of the network, including all the tangible components shown on the right. The operation of the physical layer allows the distribution of six different entangled states between the four users as shown in the quantum correlation layer (middle). The communication layer (top) exploits the entangled states to enable secure communication. The topology of the physical layer follows a hub-and-spoke model whereas the logical topology of the upper two layers is a fully connected mesh. At the network provider, a laser with a wavelength of 775 nm (green beam) is used to pump a temperature-stabilized magnesium-oxide-doped periodically poled lithium niobate crystal (MgO:ppLN) in a Sagnac-type configuration to create a polarization-entangled state ('state preparation'). The spectrum is then split into 12 International Telecommunication Union (ITU) channels (identified by different coloured symbols) by a cascade of band-pass filters ('wavelength de-multiplexing'). The resulting 12 frequency channels were combined into four single-mode fibres such that each user (Alice, Bob, Chloe and Dave) receives three frequency channels (indicated by the coloured symbols) and therefore shares a polarization-entangled pair with each of the other users ('wavelength multiplexing'). Each of the four users receives only one single-mode fibre from the network provider ('Distribution'), and analyses the polarization with a half-wave plate (HWP) and a polarizing beam splitter (PBS). The photons are then detected using one single-photon avalanche diode detector (SPAD) per user ('Polarization analysis and detection'). CH, ITU channel; DM, dichroic mirror; POLC, manual polarization controllers; YVO_4, yttrium orthovanadate plate.

without changing the source of entanglement, the type of quantum state produced or the user's hardware. At the same time, the topology can be reduced to all possible subgraphs.

The experimental set-up (Fig. 1) can be conceptualized by considering photon pairs from a polarization-entangled source that are separated into different wavelength channels (Fig. 2). Owing to energy conservation during the down-conversion process within the source, entangled photon pairs are observed only in correlated wavelength channels. Each pair of these correlated channels represents one logical link between two users (that is, a shared entangled state). Specific channels are multiplexed into a single fibre and are therefore passively rerouted to each user. Each user now receives $N-1$ channels, thus sharing a different entangled state with every other user.

To implement our network architecture using commercially available dense WDM filters, we developed a source of frequency-correlated polarization-entangled photon pairs at telecommunications wavelengths (Methods; Fig. 1). The wide spectrum of the signal and idler photons (Fig. 2) was divided into six pairs of frequency-correlated channels. Of these 12 channels, each user received three, multiplexed together in a single optical fibre. Ultimately, the source distributed six pairs of polarization-entangled photons between four different users successively, in such a way that each pair of users shares one pair of photons with each other.

To characterize the performance of the entangled-photon source, we measured the fidelity of the state produced as compared to a $|\Phi^+\rangle$ Bell state (see equation (1) in Methods). This measurement was performed directly after de-multiplexing (that is, splitting of the signal and idler photons) but before the multiplexing of several channels to each user. For this measurement, the polarization-entanglement visibility was measured in all three mutually unbiased bases just after the first cascade of band-pass filters. In this case, the fibres were compensated in only one basis (HV, where H represents the horizontal polarization and V the vertical) and the pump state of the source was changed for each measurement to compensate for the other basis. It was important to confirm that the source can provide high-quality entanglement in all available channel pairs. Our measurements show that the fidelity was greater than 97.3% for all channel pairs (Extended Data Table 1). Once we confirmed that the source of entangled photon pairs was able to provide high-quality entanglement, we connected the multiplexers and sent three channels to each of the four users.

To measure the fidelities, all 12 fibre channels were compensated in two mutually unbiased bases from the source until the measurement module to demonstrate that the entangled states were created simultaneously in all channels without further alignment. Further, the multiplexing was implemented so that three channels were detected on each of the four detectors. Entangled pairs were identified using the temporal cross-correlation functions (Fig. 3b).

We used four free-running single-photon avalanche detectors based on a passively quenched InGaAs avalanche photodiode. Three detectors operated at a detection efficiency of 2%–3% and a dark count rate of 350–1,500 Hz with a dead time of 1 μs for the measurement modules 'Bob', 'Chloe' and 'Dave'. The measurement module 'Alice' used a detector with an efficiency of about 10%, 1,000 Hz dark counts and 4 μs dead time. The rate of coincident counts varied between 10 Hz and 65 Hz for the six entangled links because the losses and detection efficiencies were unequal. We measured the visibility of all six entangled links in two mutually unbiased bases, HV and DA (where D represents diagonal polarization and A antidiagonal), and computed the fidelity. This amounts to 16 different basis settings for the HV basis and for the DA basis. Each basis setting was measured for 30 s. The count rates of the four detectors were between 21 kHz and 73 kHz. An overview over the raw counts with the polarization analysers set to H_i, where i denotes Alice, Bob, Chloe or Dave, is given in Extended Data Table 2. At this position, the maximal coincidence rate is expected.

In Fig. 3a we show the results of the Bell-state fidelity measurements. Owing to the timing uncertainty of the detectors, we are limited to a rather large coincidence window of 1 ns. As a result, detector clicks are falsely identified as pairs and deteriorate the measured fidelity. The right-hand side of Fig. 3a shows the fidelity corrected for this error; the uncorrected values are shown on the left.

Using the uncorrected fidelities and count rates (Fig. 3), we estimated a raw key rate between 10 Hz and 34 Hz, which would yield

© 2018 Springer Nature Limited. All rights reserved.

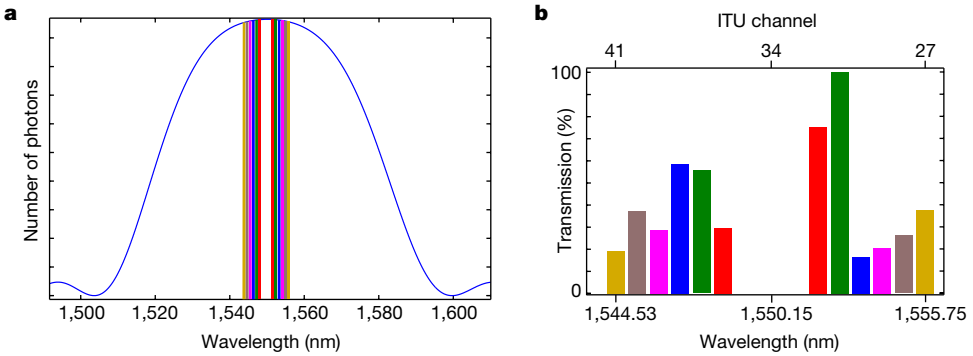


Fig. 2 | Spectrum and wavelength multiplexing. a, Spectrum of signal and idler photons; the bars are colour-coded to indicate entangled pairs of signal (lower-wavelength) and idler (higher-wavelength) photons. The spectrum of the source (blue curve) was calculated using Sellmeier equations for the MgO:ppLN used in the source[30] and information about the periodic poling from the supplier. **b**, Independently measured transmission of each of the wavelength channels (normalized).

a secure key rate between 3 Hz and 15 Hz[1,26]. A fidelity larger than 81% is necessary to obtain a positive secure key rate. Therefore, using their polarization-detection modules, the users were able to measure a non-classical polarization-correlation visibility in the HV and DA bases, from which we can calculate the lower bound on the Bell-state fidelity. These measurements show that we have successfully shared an entangled state between every pair of users.

We have successfully realized a proof-of-principle demonstration of a quantum communication network. The use of telecommunications wavelengths makes it compatible with existing infrastructure. We observed no detectable cross-talk between adjacent channels. The network architecture can be readily adapted to any other network topology. This networking concept can also be combined with previous ideas about access networks[27] and about integration into classical networks[7,17,18]. Further, our experimental demonstration of the network architecture used WDM and a polarization-entangled photon pair source. These choices are specific to the implementation and are not limited by the network architecture or logical topology. Our architecture could instead be implemented using time-division multiplexing (TDM) or time-bin entanglement. The scalability and ease of upgrading of our network architecture make it a good candidate for commercial quantum communication networks.

Our network offers all the security benefits of entanglement-based QKD and does not require trusted nodes. In contrast to networks based on active switching[20,21,25], the only limit on the communication speed in our (passive) scheme is given by the brightness of the source and the quality of the detector (efficiency, timing jitter and dead time). The finite duty cycle and switching rate characteristic of active components do not limit our network.

An alternative method to implement a fully connected quantum network with a similar topology would be to use a 1:N beam splitter and probabilistically distribute entangled photon pairs between all users. The main benefit of our wavelength-multiplexed implementation reveals itself when each user opts to de-multiplex the different wavelength channels onto m single-photon detectors (where $1 < m < N$). In this case, owing to the deterministic frequency correlations, every pair of frequency channels can be considered an independent communication link and an increase in the total key generation rate by a factor of m is achieved while maintaining the same signal-to-noise ratio of a two-party communication. Conversely, probabilistic distribution using a 1:N beam splitter would always reduce the signal-to-noise ratio as users are added.

An interesting question is how many users can be added to our network architecture while maintaining its performance. Because we used one detector per user to detect all three frequency channels, our network is linearly scalable in terms of user resources, and additional users can be added to the network without changing a user's hardware. To add a new user into a network that uses our architecture, the service provider simply multiplexes more channels into each user's fibre. As mentioned above, compared to a two-party communication scheme, detecting more than one channel on the same detector gives a higher noise level because the count rate of each detector is tripled and the coincidence rate per link is unchanged. The measured fidelities show that the network architecture is sound despite the increased noise. The number of available wavelength channels within the entangled photon spectrum and the performance of the detectors used (dark counts, timing jitter and efficiency) also limit the number of users. Our calculations (Extended Data Fig. 1) show that the network can support more

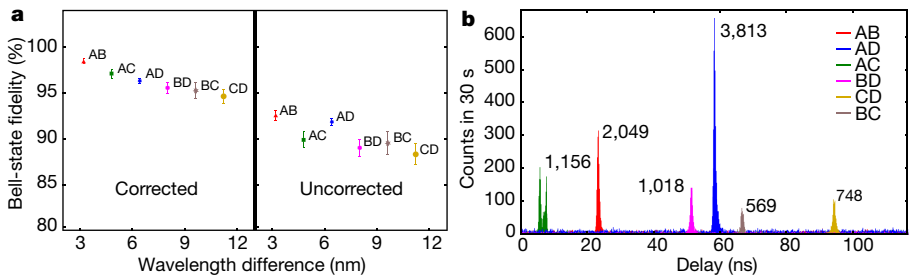


Fig. 3 | Experimental results. a, Measured Bell-state fidelities with (left) and without (right) subtraction of accidental coincidences. Each point is measured using the two WDM channels that connect the respective users (A, Alice; B, Bob; C, Chloe; D, Dave; each two-letter combination represents a link between those two users). The x axis represents the difference in wavelength between the channels of the two partner photons. The error bars correspond to one standard deviation assuming Poissonian statistics. **b**, Temporal cross-correlations between the time traces of the four users' detectors. Each cross-correlation between a pair of time traces (as indicated in the legend) has a distinct peak. The different peak positions correspond to different combinations of channel lengths and detector latencies. Because these are six different correlation functions, the unambiguous identification of the coincidence clicks is guaranteed even if some of the peaks are at the same position in time. The numbers next to each peak correspond to the total number of photon pairs, accumulated over 30 s, that arrived within 0.5 ns from the maximum of a peak at each of the two users.

© 2018 Springer Nature Limited. All rights reserved.

than eight nodes with our current detectors (1 ns coincidence window, 500 background counts per second) and more than 25 nodes when using 100 ps coincidence windows with the same noise count. However, this limitation can be avoided because all users have the option to split the signal to detect only a few or one frequency channel(s) per detector, which recovers the signal-to-noise ratio of two-party communication. Alternatively, groups of users could temporarily block frequency channels that are not currently needed for their communication. In this way, the network could also be used like an access network with switching on the user side.

Instead of continuous entanglement distribution, it is also conceivable to use a pulsed pump laser for the entangled-photon source (Methods). This would improve the signal-to-noise ratio because it would enable communication between different users to be detected in different time slots, as discussed previously[23]. As well as standard entanglement-based QKD protocols, distributed computation tasks such as the millionaire's problem[28], Byzantine fault tolerance[29] and asynchronous reference-frame agreement[29] can be implemented on our network.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0766-y.

Received: 16 February 2018; Accepted: 25 October 2018;
Published online 12 December 2018.

1. Gisin, N., Ribordy, G., Tittel, W. & Zbinden, H. Quantum cryptography. *Rev. Mod. Phys.* **74**, 145–195 (2002).
2. Stucki, D. et al. Long-term performance of the SwissQuantum quantum key distribution network in a field environment. *New J. Phys.* **13**, 123001 (2011).
3. Sasaki, M. et al. Field test of quantum key distribution in the Tokyo QKD network. *Opt. Express* **19**, 10387–10409 (2011).
4. Peev, M. et al. The SECOQC quantum key distribution network in Vienna. *New J. Phys.* **11**, 075001 (2009).
5. Xu, F. et al. Field experiment on a robust hierarchical metropolitan quantum cryptography network. *Chin. Sci. Bull.* **54**, 2991–2997 (2009).
6. Elliott, C. et al. Current status of the DARPA quantum network. *Proc SPIE* **5815**, 138–149 (2005).
7. Choi, I., Young, R. J. & Townsend, P. D. Quantum information to the home. *New J. Phys.* **13**, 063039 (2011).
8. Mao, Y. et al. Integrating quantum key distribution with classical communications in backbone fiber network. *Opt. Express* **26**, 6010–6020 (2018).
9. Patel, K. et al. Coexistence of high-bit-rate quantum key distribution and data on optical fiber. *Phys. Rev. X* **2**, 041010 (2012).
10. Korzh, B. et al. Provably secure and practical quantum key distribution over 307 km of optical fibre. *Nat. Photon.* **9**, 163–168 (2015).
11. Yin, H.-L. et al. Measurement-device-independent quantum key distribution over a 404 km optical fiber. *Phys. Rev. Lett.* **117**, 190501 (2016).
12. Ursin, R. et al. Entanglement-based quantum communication over 144 km. *Nat. Phys.* **3**, 481–486 (2007).
13. Liao, S.-K. et al. Satellite-to-ground quantum key distribution. *Nature* **549**, 43–47 (2017).
14. Takenaka, H. et al. Satellite-to-ground quantum-limited communication using a 50-kg-class microsatellite. *Nat. Photon.* **11**, 502–508 (2017).
15. Günthner, K. et al. Quantum-limited measurements of optical signals from a geostationary satellite. *Optica* **4**, 611–616 (2017).
16. Törmä, P. & Gheri, K. M. Establishing multi-party entanglement with entangled photons. *AIP Conf. Proc.* **461**, 220–228 (1999).
17. Townsend, P. D. Quantum cryptography on multiuser optical fibre networks. *Nature* **385**, 47–49 (1997).
18. Fröhlich, B. et al. A quantum access network. *Nature* **501**, 69–72 (2013).
19. Toliver, P. et al. Experimental investigation of quantum key distribution through transparent optical switch elements. *IEEE Photonics Technol. Lett.* **15**, 1669–1671 (2003).
20. Chen, T.-Y. et al. Metropolitan all-pass and inter-city quantum communication network. *Opt. Express* **18**, 27217–27225 (2010).
21. Chang, X.-Y., Deng, D.-L., Yuan, X.-X. et al. Experimental realization of an entanglement access network and secure multi-party computation. *Sci. Rep.* **6**, 29453 (2016).
22. Aktas, D. et al. Entanglement distribution over 150 km in wavelength division multiplexed channels for quantum cryptography. *Laser Photonics Rev.* **10**, 451–457 (2016).
23. Zhu, E. Y. et al. Multi-party agile QKD network with a fiber-based entangled source. In *2015 Conf. on Lasers and Electro-optics (CLEO): Science and Innovations* abstr. JW2A.10 (Optical Society of America, 2015).
24. Lim, H. C., Yoshizawa, A., Tsuchida, H. & Kikuchi, K. Broadband source of telecom-band polarization-entangled photon-pairs for wavelength-multiplexed entanglement distribution. *Opt. Express* **16**, 16052–16057 (2008).
25. Herbauts, I., Blauensteiner, B., Poppe, A., Jennewein, T. & Hübel, H. Demonstration of active routing of entanglement in a multi-user network. *Opt. Express* **21**, 29013–29024 (2013).
26. Ma, X., Fung, C.-H. & Lo, H.-K. Quantum key distribution with entangled photon sources. *Phys. Rev. A* **76**, 012307 (2007).
27. Ciurana, A. et al. Quantum metropolitan optical network based on wavelength division multiplexing. *Opt. Express* **22**, 1576–1593 (2014).
28. He, G. P. Simple quantum protocols for the millionaire problem with a semi-honest third party. *Int. J. Quant. Inf.* **11**, 1350025 (2013).
29. Islam, T. & Wehner, S. Asynchronous reference frame agreement in a quantum network. *New J. Phys.* **18**, 033018 (2016).
30. Gayer, O., Sacks, Z., Galun, E. & Arie, A. Temperature and wavelength dependent refractive index equations for MgO-doped congruent and stoichiometric $LiNbO_3$. *Appl. Phys. B* **91**, 343–348 (2008).

Acknowledgements We thank J. Slim for help with the software and E. Acuña Ortega for assistance in the laboratory. We acknowledge financial support from the Austrian Research Promotion Agency (FFG) Projects—Agentur für Luft- und Raumfahrt (FFG-ALR contracts 854022 and 866025), the European Union (EU) under Horizon 2020 contract number FETOPEN-801060 quantum-enhanced on-chip interference microscopy (Q-MIC) and the Austrian Academy of Sciences.

Reviewer information *Nature* thanks V. Martin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The set-up was built by S.W. and the experiment was conducted by S.W. and S.K.J. The network architecture was conceived by S.K.J. and S.W. The source was designed by F.S., S.W. and S.K.J. H.H. helped with the detection of the single photons. R.U. contributed to the experimental design, source and network and to supervising the project. The paper was written by S.W., S.K.J. and F.S. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0766-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.W. or R.U.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Polarization-entangled photon source. The experiment consisted of a source of bipartite polarization-entangled photon pairs, multiplexing and de-multiplexing modules, and user hardware. The source was based on type-0 spontaneous parametric down-conversion in a 4 cm-long magnesium-oxide-doped periodically poled lithium niobate (MgO:ppLN) bulk crystal with a poling period of 19.2 μm. The type-0 process converts, with low probability, one pump photon with a wavelength of 775.075 nm from a continuous-wave laser to a co-polarized signal and idler photons in the telecommunications C-band[31–33].

The MgO:ppLN crystal was bi-directionally pumped inside a Sagnac-type set-up (Fig. 1)[24,34,35], creating a polarization-entangled state in two wavelength channels:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|V_{\lambda_1}V_{\lambda_2}\rangle + |H_{\lambda_1}H_{\lambda_2}\rangle) \quad (1)$$

The spatial mode that contains the signal and idler photons from the source was coupled into one single-mode fibre and spectrally split by a cascade of band-pass filters. The spectrum of the signal and idler photons was centred at 1,550.15 nm (Fig. 2) and the filters were chosen to be symmetric with respect to this centre wavelength. We used 100 GHz band-pass filters as defined by the ITU in G.694.1. On the red side of the spectrum we used ITU frequency channels 27–32; we used channels 36–41 on the blue side. Owing to the well-defined pump wavelength of the continuous-wave laser and energy conservation during down-conversion, we obtained polarization entanglement between pairs of channels (27 and 41, 28 and 40, and so on). Each user receives three channels (Fig. 2) via one fibre and used a polarization analysis module to measure in the HV or DA polarization basis. Single-photon detection events were time-tagged and two-photon coincidence events were identified within a coincidence window of 1 ns. Fibre polarization controllers were used to neutralize the birefringence of the optical fibres.

The energy correlations of the signal and idler photons (see equation (1) and Fig. 2) produced by the source were used to separate these modes into separate fibres (de-multiplexing). We used channels 27 to 32 for the signal photons, while the idler photons were collected in channels 36 to 41. The corresponding wavelengths are provided in Extended Data Table 1. Entangled photon pairs are found in pairs of channels that have the same spectral distance from the centre wavelength. This means that the channel pairs 27 and 41, 28 and 40, and so on, each share a polarization-entangled state (Fig. 2). A viable alternative to the cascade of dense WDM filters is an arrayed waveguide grating, provided that the polarization-dependent loss is low enough.

The 12 channels were combined into four fibres using two band-pass filters per fibre, so that three channels reach each one of the four users via one fibre. This way, every pair of users shares a pair of channels and therefore entangled photons (Fig. 1). The three channels received by each user were analysed on a single polarization analysis module with a single photon detector attached. Each user implemented a basis choice by rotating a half-wave plate.

QKD and signal-to-noise ratio considerations. In general, the chief drawback of our architecture is the amount of noise introduced by detecting several channels on a single detector. The noise results in a loss of fidelity (or quality of the entanglement). To implement a QKD scheme, all users would announce their time tags and correlation functions (Fig. 3b) publicly, so that everybody is able to ignore counts that do not belong to their communication and therefore improve the signal-to-noise ratio. The security of the implementation is preserved, because the annunciation of the time tags does not contain any information about the basis choice or the outcome of the measurement. This improvement is related to the total losses in the system, as shown in Extended Data Fig. 1. This is equivalent to ignoring all global ($n > 2$)-fold events. However, it makes a noticeable effect only for very low-loss scenarios, as can be seen from Extended Data Fig. 1.

In other words, the count rate S_i per user in a network with N nodes, the link and system efficiency η and the dark count rate D would be reduced by a term that scales proportional to the coincidence probability:

$$S_i = D + (N-1)\frac{P}{2}\eta - (N-2)\frac{P}{2}\eta^2$$

with P being the total number of available pairs in the spatial and spectral collection mode of the source. The rate of coincidence clicks can be estimated as

$$C = \frac{P}{2}\eta^2 + \tau S_i^2$$

The accidental coincidences (τS_i^2) account for the minimum number of coincidences observable using a coincidence window of length τ and therefore reduce the contrast.

Another substantial improvement can be made by decreasing the coincidence time window. This can be achieved by using faster detectors with a much smaller timing jitter. For example, reducing the coincidence window to 100 ps results in the maximum fidelities shown in Extended Data Fig. 1.

Scalability. Our network architecture is easily scalable and users can be added and removed without any change to the user's hardware. However, like most existing network hardware there are limitations to the scalability. The three main limitations are: first, the brightness of the source; second, the limited bandwidth of the source, which dictates how many wavelength channels can be used; and third, accidental coincidences, which contribute substantially to the quantum-bit error rate and increase markedly with the number of users. The first limitation can be overcome by using more or longer waveguides and crystals, and stronger pumping; the second can be overcome by using narrower wavelength channels; and the third can be mitigated by using the method described above to help reduce the noise. Naturally, using faster detectors and therefore shorter coincidence windows can also help. A pulsed pump experiment would further mitigate the problem of accidental coincidences by defining fixed time slots for the arrival of each channel at the detector.

Our network architecture offers the advantage of simultaneous communication between one node and every other node. Nevertheless, should one user choose to completely block the signal from a set of other users, an active switch capable of selecting certain channels can be used. This would allow users to control the network topology and create custom subgraphs without the intervention of the service provider. Further, detecting only a chosen subset of channels would limit the accidental coincidence rates and allow for faster communication with a chosen subgraph.

Pulsed network scheme. The drawback of the scheme presented here is the increase in the accidental count rates due to the multiplexing of many quantum channels onto a single detector. This limitation can be completely overcome by using a pulsed scheme. Consider the experiment presented here, but using a pulsed laser with a pulse width much smaller than the detector jitter. Further, each of the N users has gated detector(s) for which the gate is opened $N-1$ times for each laser pulse. Each opening of the gate corresponds to the time delay between different coincidence peaks among all users with each user in question. With ideal detectors, the performance of the pulsed scheme will be equivalent to $N-1$ separate quantum communication set-ups with the same detectors and comparable count rates per link. When using real-world detectors such as InGaAs SPADs that have a large dead time, the performance of our pulsed network scheme can exceed that of $N-1$ independent set-ups. When the dead time of the detector is larger than the interval between opening each of the $N-1$ gates for each pulse of the source, a noise count in one gate prevents the occurrence of a noise count in all subsequent gates within the dead time. This suppression of noise clicks can lead to improved key rates and quantum-bit error rate[1]. The advantage is strongest when there is only one photon pair in the given set of $N-1$ links per user.

This pulsed network scheme would require an additional gating signal to be sent to each user. Further, it could be unsuitable for mobile nodes because all nodes need to compensate the delays to all other nodes. However, for fixed users, the pulsed network scheme is ideal and greatly improves the network throughput by reducing the accidental count rate by a factor equal to the duty cycle of the gating.

Multiplexing and types of entanglement. The logical network topology that we have outlined here is independent of the type of entanglement or multiplexing used. Nevertheless, different types of multiplexing have advantages. For example, a scheme based on WDM has a few advantages over that based on TDM. First, the active switching used in TDM is prone to mechanical breakdown and in more complex networks several switches may need to operate synchronously. Second, a bright source can produce multiple photon pairs within a single coincidence time window. However, the probability that multiple pairs are produced in exactly the same wavelength channel is negligible. Thus, WDM-based networks could have a distinct advantage. Third, introducing an additional TDM channel will reduce the coincidence rates seen by all users, but an additional WDM channel will not affect existing connections. Last, cross-talk between the channels is not harmful, because photons in the wrong channel would, owing to the different delay introduced by the WDM filters, contribute to only the accidental rate and not be seen as a separate coincidence peak. On the other hand, a TDM-based scheme would need only $2N$ channels. Large-scale networks could also combine the advantages of WDM and TDM by using both together.

Fibre-based quantum communication has often been performed using time-bin entanglement[22,36] to avoid having to compensate for the birefringence of the fibre. However, this requires the service provider and users to have matched and stabilized interferometers (the stabilization of which often requires another stable laser). Although the logical network topology is compatible with this form of entanglement, we chose to use polarization entanglement because it simplifies the user's hardware. Changes in the birefringence of the optical fibre are easily monitored and compensated for by using regular test signals.

© 2018 Springer Nature Limited. All rights reserved.

In principle, our network architecture is not limited to the use of single photons. It is also conceivable to perform continuous-variable QKD with a source of entanglement, as proposed previously[37].

Usage scenarios. Quantum communication is often thought of as a purely academic concept or experiment. However, the technology is mature enough to consider practical problems regarding the deployment and use of QKD links. Typical classical networks consist of smaller local-area networks (LANs) and similar as well as much larger inter-city networks. Both a LAN and an inter-city network have a limited number of users. To connect a large number of users together the networks must be interconnected to create the 'internet'. Similarly, a quantum internet must also be an interconnection of several networks. A single user in our network architecture could be replaced by a quantum repeater or entanglement swapping set-up to interconnect several similar quantum networks. The most substantial differences between the two types of network are the distances spanned, the costs and the target market.

To realize a cheap LAN with current technologies and our network architecture, we propose using a cheaper type of single photon detector—SPADs. These typically have a low detection efficiency and large timing jitter. As can be seen by extrapolating Extended Data Fig. 1a, the network will be able to tolerate more than 30 dB of loss with up to 12 users. This loss is more than sufficient to account for a few kilometres of optical fibre, the heralding efficiency of a typical source and the poor detection efficiency of SPADs.

Inter-city networks naturally cost much more than a LAN. Our network architecture can be used to build large-scale inter-city quantum networks by using high-efficiency low-timing-jitter detectors such as nano-wire detectors. Extended Data Fig. 1b shows that we can tolerate more than 43 dB of loss with up to 25 users spanning distances of more than 200 km.

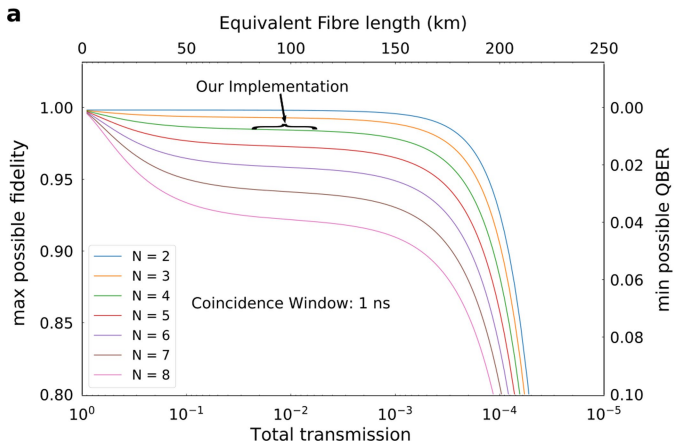
Further, in any network it is always advantageous to make the user's hardware requirements as simple as possible, with the centralized network hardware having the majority of the complexity. We have designed our network architecture along these principles, with almost all complexity in the three centralized stages—source, de-multiplexing and multiplexing.

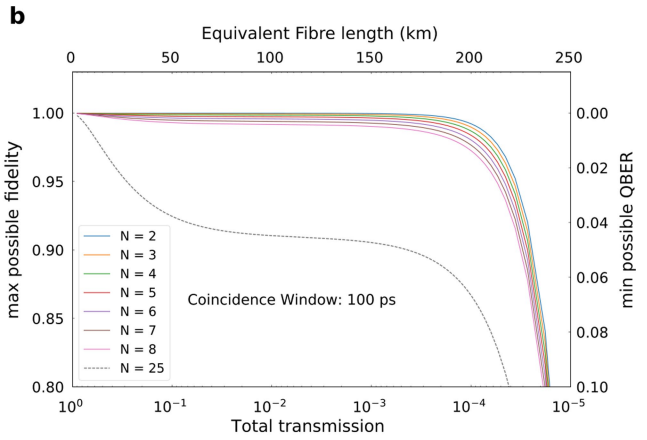
Data availability

The data that support the findings of this study are available from the corresponding authors on request.

31. Oh, J., Antonelli, C. & Brodsky, M. Coincidence rates for photon pairs in WDM environment. *J. Lightwave Technol.* **29**, 324–329 (2011).
32. Ghalbouni, J., Agha, I., Frey, R., Diamanti, E. & Zaquine, I. Experimental wavelength-division-multiplexed photon-pair distribution. *Opt. Lett.* **38**, 34–36 (2013).
33. Monteiro, F., Martin, A., Sanguinetti, B., Zbinden, H. & Thew, R. T. Narrowband photon pair source for quantum networks. *Opt. Express* **22**, 4371–4378 (2014).
34. Kim, T., Fiorentino, M. & Wong, F. N. C. Phase-stable source of polarization-entangled photons using a polarization Sagnac interferometer. *Phys. Rev. A* **73**, 012316 (2006).
35. Fedrizzi, A., Herbst, T., Poppe, A., Jennewein, T. & Zeilinger, A. A wavelength-tunable fiber-coupled source of narrowband entangled photons. *Opt. Express* **15**, 15377–15386 (2007).
36. Inagaki, T. et al. Entanglement distribution over 300 km of fiber. *Opt. Express* **21**, 23241–23249 (2013).
37. Roslund, J., De Araújo, R. M., Jiang, S., Fabre, C. & Treps, N. Wavelength-multiplexed quantum networks with ultrafast frequency combs. *Nat. Photon.* **8**, 109–112 (2014).

© 2018 Springer Nature Limited. All rights reserved.





Extended Data Fig. 1 | Calculated fidelities and quantum-bit error rate (QBER) for two to nine users versus the system efficiency and equivalent fibre length, assuming an attenuation of 0.2 dB km^{-1}. **a**, Using detectors with a 1 ns timing jitter. This is great for cheap networks with low losses (those over a small area such as a LAN). **b**, Using detectors with a 100 ps jitter allows us to sustain much higher losses and many more users. This is useful for long-distance inter-city links. Both graphs were calculated using a generated pair rate of 1.7 million pairs per second and a dark count rate of 500 per second per detector.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Measured fidelities

ITU Ch. Numbers	Channel Wavelengths (nm)	Fidelity (±0.3%)
27 / 41	1555.75 / 1544.53	98.0 %
28 / 40	1554.94 / 1545.32	98.7 %
29 / 39	1554.13 / 1546.12	99.1 %
30 / 38	1553.33 / 1546.92	99.0 %
31 / 37	1552.52 / 1547.72	99.2 %
32 / 36	1551.72 / 1548.52	97.3 %

The Bell-state fidelity of the entangled state produced by the source is measured directly at the channel pairs before multiplexing.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Count rates

	Alice	Bob	Chloe	Dave
Alice	2204203	2049	1156	3813
Bob		878692	569	1018
Chloe			636268	748
Dave				1231478

Measured coincidence counts between two users in 30 s are given for all four measurement stations at the setting HHHH. The total counts in 30 s at each station are shown on the diagonal.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0764-0

Experimental realization of on-chip topological nanoelectromechanical metamaterials

Jinwoong Cha[1,2], Kun Woo Kim[3] & Chiara Daraio[2]*

Guiding waves through a stable physical channel is essential for reliable information transport. However, energy transport in high-frequency mechanical systems, such as in signal-processing applications[1], is particularly sensitive to defects and sharp turns because of back-scattering and losses[2]. Topological phenomena in condensed matter systems have shown immunity to defects and unidirectional energy propagation[3]. Topological mechanical metamaterials translate these properties into classical systems for efficient phononic energy transport. Acoustic and mechanical topological metamaterials have so far been realized only in large-scale systems, such as arrays of pendulums[4], gyroscopic lattices[5,6], structured plates[7,8] and arrays of rods, cans and other structures acting as acoustic scatterers[9–12]. To fulfil their potential in device applications, mechanical topological systems need to be scaled to the on-chip level for high-frequency transport[13–15]. Here we report the experimental realization of topological nanoelectromechanical metamaterials, consisting of two-dimensional arrays of free-standing silicon nitride nanomembranes that operate at high frequencies (10–20 megahertz). We experimentally demonstrate the presence of edge states, and characterize their localization and Dirac-cone-like frequency dispersion. Our topological waveguides are also robust to waveguide distortions and pseudospin-dependent transport. The on-chip integrated acoustic components realized here could be used in unidirectional waveguides and compact delay lines for high-frequency signal-processing applications.

Nanoelectromechanical systems[16,17] can be employed to build on-chip topological acoustic devices, thanks to their ability to transduce electrical signals into mechanical motion, which is essential in applications. Moreover, nonlinear dynamic phenomena are easily accessible in nanoelectromechanical devices. For example, previous

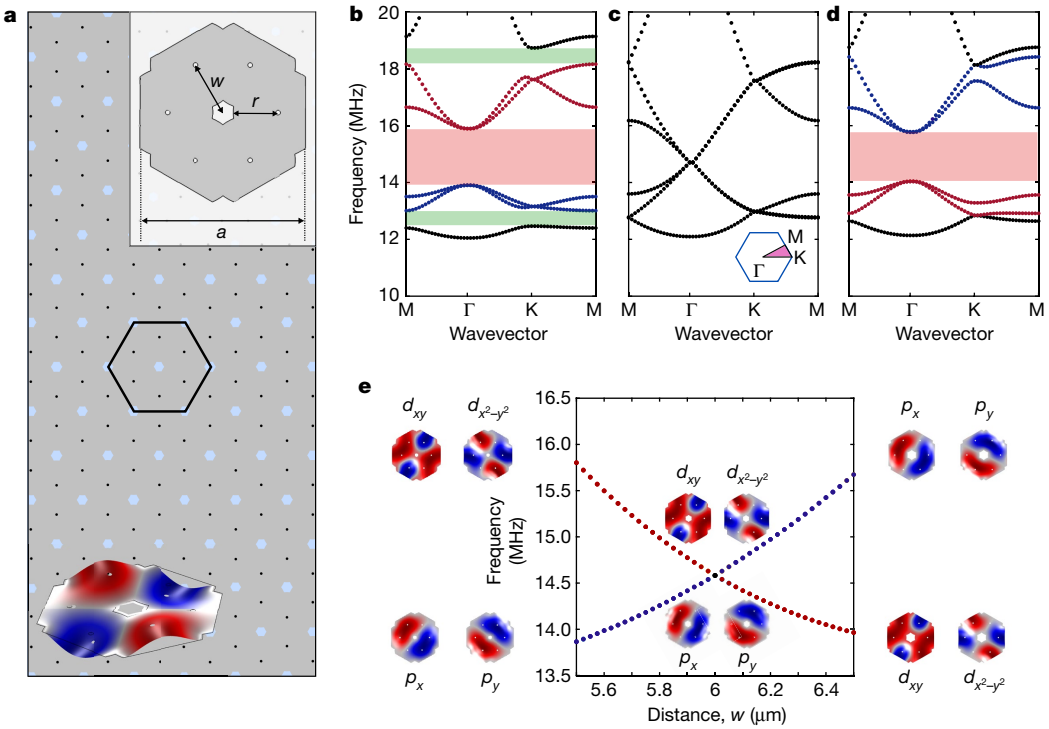


Fig. 1 | Unit cell geometry and topological phase transitions. a, Schematic of a two-dimensional NEMM. The grey area represents the SiN nanomembrane suspended over a highly doped n-type silicon substrate. The black dots, forming a honeycomb lattice, represent etch holes. The light-blue hexagons represent the unetched thermal oxide, acting as fixed boundaries. The unit cell geometry (black solid hexagon) is shown in the right inset, with relevant parameters. An example flexural mode is shown in the left inset. The topological phases are controlled by changing the centre-to-hole distance w. r denotes the radius of the etching path from the etch holes. **b**–**d**, Frequency dispersion curves along a boundary of the irreducible Brillouin zone MΓKM, for $w = 5.5\,\mu$m (**b**), $6.0\,\mu$m (**c**) and $6.5\,\mu$m (**d**); $r = 4.9\,\mu$m. The red- and green-shaded regions correspond to topological and non-topological bandgaps, respectively. **e**, Eigenfrequencies above and below the topological bandgap at the Γ point, as a function of w. Blue (red) dots denote the eigenfrequencies for flexural modes p_x and p_y (d_{xy} and $d_{x^2-y^2}$). The flexural mode shapes are presented for $w = 5.5\,\mu$m (left), $6.0\,\mu$m (middle) and $6.5\,\mu$m (right).

[1]Department of Mechanical and Process Engineering, ETH Zurich, Zurich, Switzerland. [2]Engineering and Applied Science, California Institute of Technology, Pasadena, CA, USA. [3]Korea Institute for Advanced Study, Seoul, South Korea. *e-mail: daraio@caltech.edu

© 2018 Springer Nature Limited. All rights reserved.

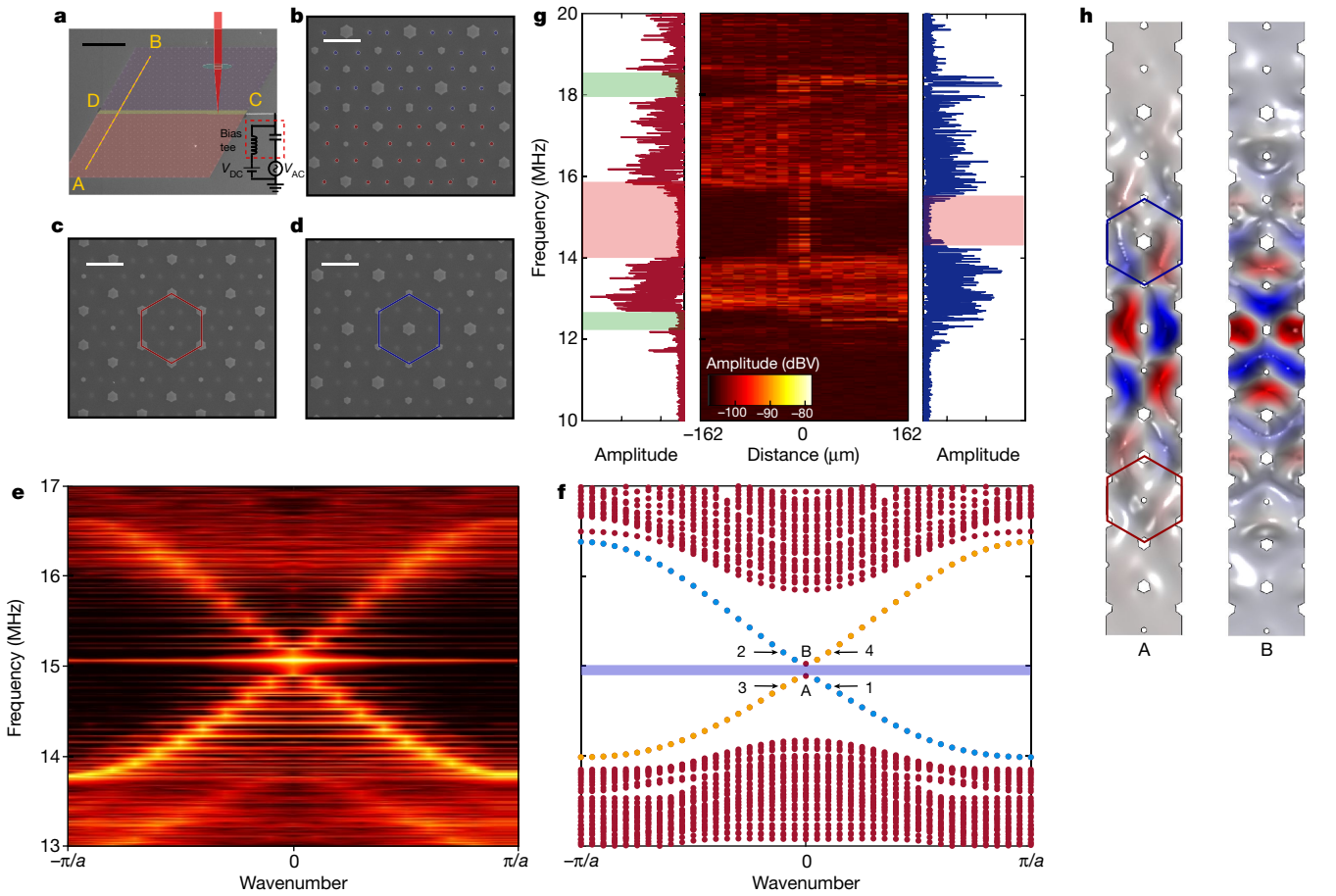


Fig. 2 | Characterization of topological edge states. a, Scanning electron microscope (SEM) image of a straight topological edge waveguide. The two different topological phases are fasle-colour-shaded in blue (non-trivial) and red (trivial). Flexural membrane motions are excited by simultaneously applying a constant and alternating voltages ($V_{DC} = 2$ V, $V_{AC} = 20$ mV) to the excitation electrode via a bias tee. Scale bar, 100 μm. **b**–**d**, SEM images of an edge region (**b**; the yellow-shaded strip C–D in **a**) a trivial lattice with $w = 5.5$ μm (**c**; the red-shaded area in **a**) and a non-trivial lattice with $w = 6.5$ μm (**d**; the blue-shaded area in **a**). Scale bars, 10 μm. The red and blue dots in **b** denote the lattice points for $w = 5.5$ μm and $w = 6.5$ μm, respectively. The red (**c**) and blue (**d**) hexagons represent the unit cells for $w = 5.5$ μm and $w = 6.5$ μm, respectively. **e**, **f**, Experimental (**e**) and numerical (**f**) frequency dispersion curves along the edge waveguide (C–D in **a**). Yellow and light-blue dots in the edge-state dispersion in **f** represent propagating waves for two opposite pseudospins. Time evolutions of the mode shapes at points 1, 2, 3 and 4 are provided in Supplementary Videos 1–4. **g**, Frequency responses for 19 different sites along the yellow dashed line A–B in **a** (middle panel). The left and right panels represent frequency responses at sites A and B, respectively. The red- and green-shaded regions represent the bandgaps. **h**, Flexural modes for points A and B in the dispersion shown in **f**. The width of the strip is 18 μm, identical to the lattice parameter *a*.

studies of systems with few degrees of freedom have demonstrated quantum-analogous phenomena, like cooling and amplification[18], and Rabi oscillation[19,20]. One-dimensional nanoelectromechanical lattices are a different class of nanoelectromechanical devices used to study lattice dynamics, for example, in waveguiding[21–23] and energy focusing[24]. Recently, one-dimensional nanoelectromechanical lattices made of SiN nanomembranes have demonstrated active manipulation of phononic dispersion, leveraging electrostatic softening effects and nonlinearity[23].

To design our topological nanoelectromechanical metamaterial (NEMM), we implemented the well known extended honeycomb lattice scheme[25]. The approach emulates electronic topological insulators for bosonic excitations in metamaterials. The extended honeycomb lattice contains six sites in a unit cell, satisfying C_6 crystalline symmetry[25]. This is an effective design strategy for device applications because of its geometrical simplicity. This lattice exploits Brillouin-zone folding to demonstrate a double-Dirac cone at the Γ point of the Brillouin zone. This zone-folding method has recently been used in various topological elastic[8,15], acoustic[9,11] and photonic[25,26] systems, by introducing the concept of pseudospins that satisfy Kramers theorem[3]. Brillouin-zone folding allows us to realize a pseudotime-reversal-symmetry-invariant system, where an anti-unitary, pseudotime-reversal operator U_T (where $U_T U_T^\dagger = -I$) is defined from the crystalline symmetry (C_6) of the lattice[25]. Despite the practicality of the structure, the consequent topological edge states are robust only against defects that preserve local C_6 symmetry, which is an inevitable drawback of crystalline-symmetry-based designs.

We realize these topological properties in our NEMM by periodically arranging etch holes, of diameter 500 nm, in an extended honeycomb lattice (Fig. 1a). The etch holes enable a buffered oxide etchant to partially remove the sacrificial thermal oxide layer and release the SiN suspended membranes (Extended Data Fig. 1). We engineer the topological phases of the lattice by changing the distance between etch holes, *w* (Fig. 1a). Our NEMM consequently forms a flexural phononic crystal, consisting of a periodic array of free-standing SiN nanomembranes. The average thickness of the nanomembranes is about 79 nm. The average vacuum gap distance between the SiN layer and the highly doped silicon substrate is about 147 nm. These values are estimated considering the partial etching rate of the SiN in the buffered oxide etchant (about 0.3 nm min^{-1}).

We perform finite element simulations using COMSOL Multiphysics (https://www.comsol.com/), to numerically compute frequency dispersion curves for a unit cell with a lattice parameter $a = 18$ μm (Extended Data Fig. 2). We vary the distance between two neighbouring holes, *w*, from 5.5 μm to 6.5 μm (Fig. 1b, e). For a unit cell with $w = 6.0$ μm $= a/3$, a double Dirac cone is present around 14.55 MHz at the Γ point of the Brillouin zone (Fig. 1c). The frequency dispersion curves typically start

© 2018 Springer Nature Limited. All rights reserved.

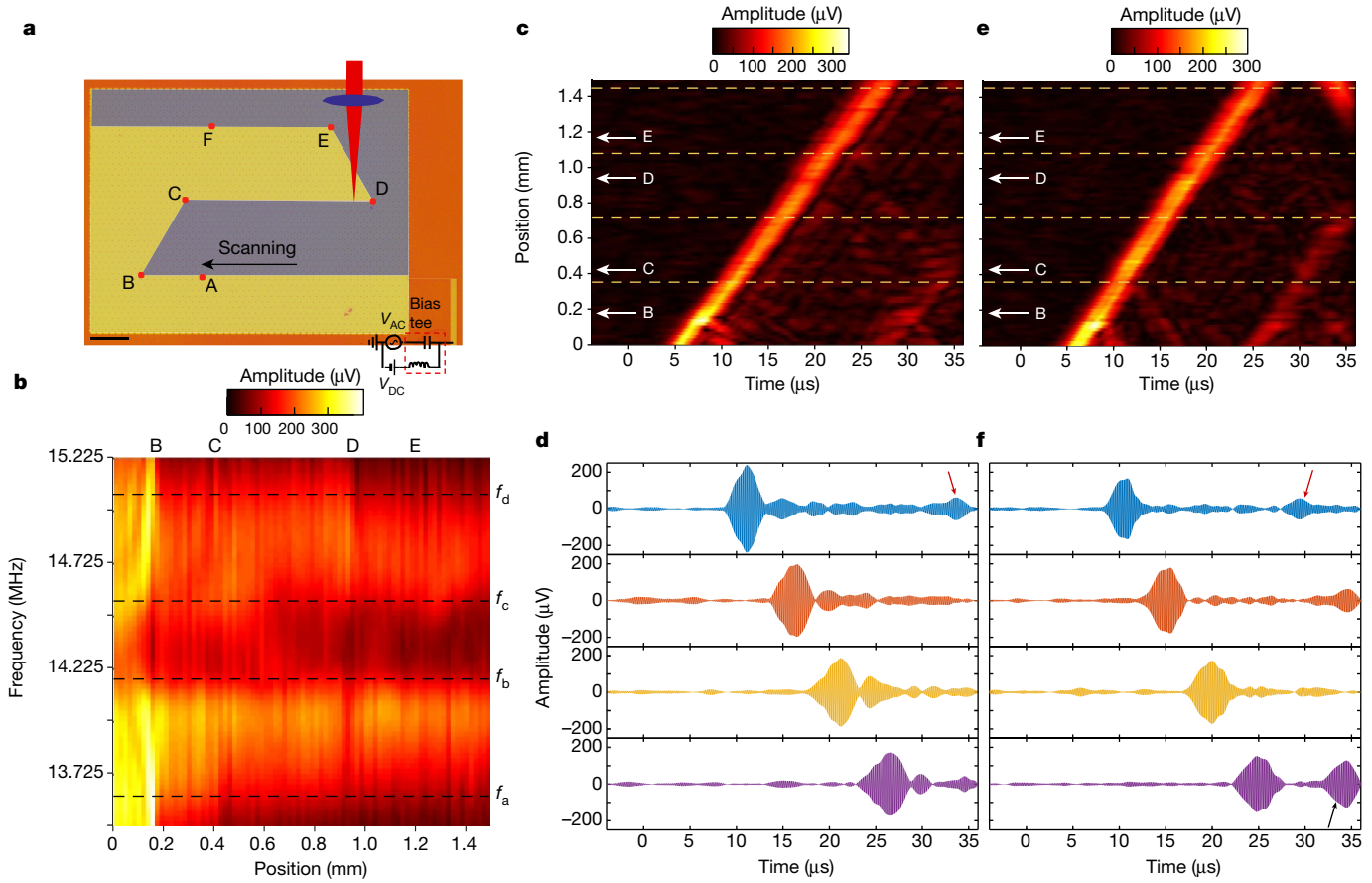


Fig. 3 | Waveguide robustness against imperfections. a, Optical microscope image of a zigzag topological edge waveguide. The blue- and yellow-shaded regions represent topologically trivial and non-trivial lattices, respectively. Time-domain responses are measured along the edge waveguide from points A to F. Points B, C, D and E denote the corners. The flexural membrane motions are excited by simultaneously applying a constant and a chirped voltage signal with frequencies ranging from 12.8 MHz to 15.8 MHz. The applied voltages are $V_{DC} = 15$ V and $V_P = 30$ mV, where V_P is the amplitude of the chirped signal. Scale bar, 100 μm. **b**, The colour scale represents the amplitudes from wavelet analyses at different positions of the edge waveguide. The frequencies ranging from f_a to f_b and from f_c to f_d represent propagating edge states. **c**, **e**, Spatiotemporal responses along the edge waveguide in a space–time domain. Pulses with centre frequencies of 13.85 MHz (**c**) and 14.75 MHz (**e**) and a bandwidth of 0.3 MHz are considered. **d**, **f**, Time-domain responses for pulses with centre frequencies of 13.85 MHz (**d**) and 14.75 MHz (**f**) at the 20th (blue), 40th (red), 60th (yellow) and 80th (purple) unit cells along the edge waveguide, which are highlighted with yellow-dashed lines in **c** and **e**. The red (black) arrow indicates the reflected pulse from the input (output) boundary.

from around 12 MHz, because of the presence of clamped boundaries. The frequency dispersion curves for $w = 5.5$ μm and 6.5 μm show the emergence of approximately 1.8-MHz-wide band gaps at the Γ point, ranging from 14 MHz to 15.8 MHz. The lattice with $w = 5.5$ μm exhibits two additional bandgaps below and above the centre bandgap around 15 MHz (Fig. 1b), while the lattice with $w = 6.5$ μm (Fig. 1d) does not. The four vibrational modes, p_x, p_y, d_{xy} and $d_{x^2-y^2}$, at the Γ point are degenerate at the Dirac point for the lattice with $w = 6$ μm (Fig. 1c, e). The four degenerate modes are split into two separate degenerate modes, for $w < 6$ μm and $w > 6$ μm (Fig. 1e), opening a bandgap. The band inversion between the dipole vibrational modes (p_x, p_y) and the quadrupole ones (d_{xy}, $d_{x^2-y^2}$) appears at the Γ point for $w > 6$ μm, which supports the topological non-triviality of the lattice. To confirm the presence of the pseudospins, we derive a Hamiltonian matrix for pseudospin states $p_\pm = p_x \pm ip_y$ and $d_\pm = d_{x^2-y^2} \pm id_{xy}$ around the Γ point (Methods). We apply the $\boldsymbol{k}\cdot\boldsymbol{p}$ perturbation method (see Methods) to the wave equation for a thin plate, $D\nabla^4 W = -\rho h(\partial^2 W/\partial t^2)$, and show the similarity with the Bernevig–Hughes–Zhang model for CdTe/HgTe/CdTe quantum wells[27].

To investigate the topological properties of our NEMMs experimentally, we fabricate a straight topological edge waveguide (Fig. 2a, b), formed at the interface of the topologically trivial ($w = 5.5$ μm, Fig. 2c) and non-trivial ($w = 6.5$ μm, Fig. 2d) lattices. Topological edge states do not exist at free boundaries of our systems, owing to the lack of C_6 symmetry. The number of unit cells of each phase is approximately 200, so that the edge waveguide has 20 supercells with 18-μm one-dimensional lattice spacing. To characterize the edge states, we excite the flexural motion of the membranes by applying a dynamic electrostatic force, $F \propto (V_{DC} + V_{AC})^2$, to the excitation electrode. Here, V_{DC} and V_{AC} are the constant and alternating voltages, which are simultaneously applied between the excitation electrode and the grounded substrate (Fig. 2a). We perform measurements using a home-built Michelson interferometer with a balanced homodyne detection scheme (Methods). To obtain the dispersion curves of the edge states, we measure the frequency responses of 20 sites along the edge waveguide, by spatially scanning the measurement points (yellow strip, Fig. 2a) with 18-μm steps (Extended Data Fig. 3). The Dirac-like edge-state frequency dispersion curves, isolated from the bulk dispersion, are present in the frequency range 14.1–15.8 MHz, showing good agreement with the numerical dispersion curves (Fig. 2f). We also observe a defect mode at the crossing point of the edge-state dispersion curves (Fig. 2e). This stems from a point-defect mode at the boundary near the excitation region (Extended data Fig. 4a). The broken C_6 symmetry at the interface (Fig. 2b) induces a small bandgap in the middle of the edge-state dispersions (Fig. 2e, f). Despite the presence of the bandgap, the defect mode is allowed to transmit non-negligible energy to the end of the waveguide owing to the long decay length of the evanescent mode (Extended data Fig. 4b).

We also characterize the localization of the edge states, by scanning the measurement point across the waveguide (yellow dashed line A–B

© 2018 Springer Nature Limited. All rights reserved.

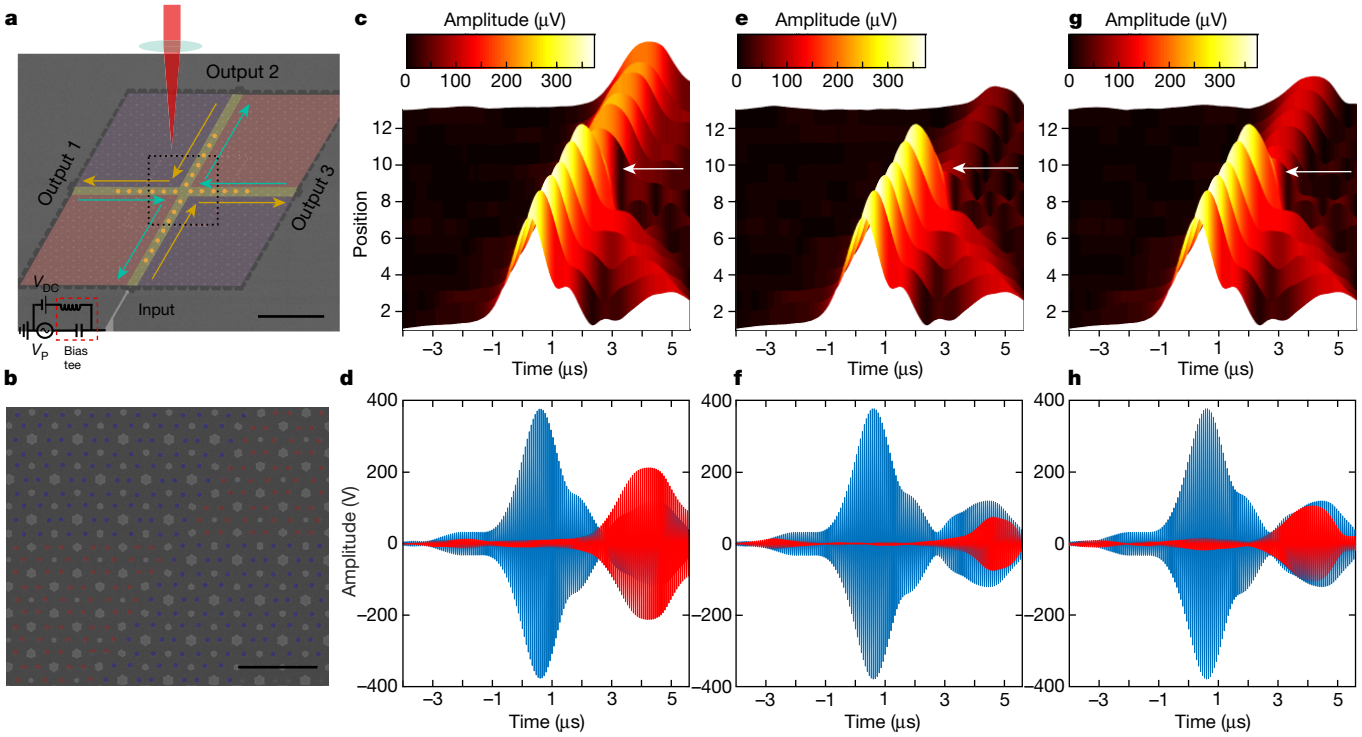


Fig. 4 | Pseudospin-dependent wave propagation. a, SEM image of a pseudospin-filter configuration. Scale bar, 100 μm. The two different topological phases are false-colour-shaded in red (trivial, $w = 5.5$ μm) and blue (non-trivial, $w = 6.5$ μm). Flexural motions are excited from the electrode ($V_{DC} = 15$ V, $V_P = 22.5$ mV). The yellow and cyan arrows represent propagating directions for different pseudospin states. **b**, Close-up view of the region marked by the dotted square in **a**. The red and blue dots denote the lattice points for trivial and non-trivial phases. Scale bar, 30 μm. **c**, **e**, **g**, Envelopes of propagating pulses (15.2-MHz centre frequency, 0.5-MHz bandwidth) in the space–time domain. The position represents the measurement points along the edge waveguides. **c**, Input to output port 1; **e**, input to output port 2; and **g**, input to output port 3. The crossing points are indicated by white arrows. **d**, **f**, **h**, Time-domain responses of the propagating pulses at position 1, input side (blue) and position 13, output side (red). **d**, Output port 1; **f**, output port 2; and **h**, output port 3.

in Fig. 2a), also with 18-μm steps. The edge states are strongly localized (Fig. 2h) within ±36 μm from the interface (Fig. 2g). Beyond this range, the frequency responses (Fig. 2g) start to show clear bandgaps, with frequency ranges and widths similar to the numerical dispersion relations shown in Fig. 1b and d. The trivial lattice presents three bandgaps (Fig. 2g, left) and the non-trivial lattice shows only one topological bandgap (Fig. 2g, right), as predicted by the numerical frequency dispersion (Fig. 1b–d). The frequency responses show evidence of different topological phases in the two lattices, with $w = 5.5$ μm and 6.5 μm, confirming that the waveguiding effect is topological.

One remarkable feature of topological edge modes is their robustness to waveguide imperfections, such as sharp corners. To study this, we fabricate a long, distorted edge waveguide that includes four corners with two 60° and two 120° angles (Fig. 3a). This waveguide consists of 134 unit cells. We perform steady-state measurements (Methods) and confirm the presence of the topological edge states (Extended Data Fig. 5a). The bandgap of the trivial phase is observed in the range 13.7–15.1 MHz and that for the non-trivial phase in the range 13.7–14.8 MHz. To confirm immunity to back-scattering, we measure transient responses of propagating pulses over 84 unit cells along the edge waveguide (points A to F in Fig. 3a). We excite chirped signals with frequencies ranging from 12.8 MHz to 15.8 MHz (Methods). The propagating pulses that correspond to the edge states (frequencies of 13.7–14.1 MHz and 14.5–14.9 MHz) exhibit small-amplitude decays despite the presence of the corners (Fig. 3b). To closely examine the propagating pulses, we select two pulses whose centre frequencies (13.85 MHz in Fig. 3c and 14.75 MHz in Fig. 3e) lie in the lower and upper edge states with respect to the small bandgap around 14.3 MHz. The propagation speeds extracted from the data are 69.1 m s^{-1} (for the 13.85-MHz pulse) and 78.6 m s^{-1} (for the 14.75-MHz pulse). We note that the waves propagate along the entire length of the edge waveguide without visible leakage in the bulk from the excitation point (Fig. 3c, e) and with little backscattering from the corners (Fig. 3c–f). This confirms that the energy transport is very stable and strongly confined at the interface. The low backscattering and signal decay may be attributable to scattering of the parasitic bulk modes (which coexist with the edge modes) as well as to partial pseudospin-mode conversion due to imperfect C_6 symmetry at the corners. The ability to introduce sharp corners (as in B, C, D and E) allows longer waveguides (for example, in delay lines) to be designed within the same device size.

Another crucial aspect of topological insulators is unidirectional propagation for distinct pseudospin modes. To characterize this, we fabricate another NEMM with a spin-splitter configuration consisting of four domain walls that has been employed in several previous studies[11,12] (Fig. 4a, b). Such geometry allows us to use a simpler pseudospin selective excitation. We send voltage pulses to the excitation electrode and measure transient responses of the propagating pulses (Methods). We scan 13 sites (7 sites from the input channel and 6 sites from each output channel) near the crossing point of the channels (Fig. 4b). Note that the steady-state frequency responses at the end of the three output ports (Fig. 4a) exhibit almost identical edge-state responses owing to boundary scattering (Extended Data Fig. 6b–d). The pulse we investigate has a 15.1-MHz centre frequency and 0.5-MHz bandwidth, which is enough to cover the broad frequency ranges of edge states. In this configuration, the propagating direction of a pseudospin state depends on the spatial configuration of the two topological phases, $w = 5.5$ μm and 6.5 μm (Fig. 4a). The pseudospin states are filtered to have a single dominant state in the input port (yellow arrow in Fig. 4a). After the signal passes the input channel, the filtered spin state mainly propagates to output port 1 and 3 (yellow arrows in Fig. 4a) as shown in Fig. 4c, d, g and h. The edge state leading to port 2 supports pseudospin modes that are opposite to the input modes in the propagation direction (cyan arrow, Fig. 4a). As such, we would expect no signal to reach port 2. However, we observe a small, but visible energy propagation (Fig. 4e, f).

© 2018 Springer Nature Limited. All rights reserved.

This unexpected penetration might arise from the partial conversion of pseudospin modes at the centre crossing point, where the C_6 symmetry is broken. Nonetheless, the results confirm that the propagation direction depends on the type of pseudospin. The use of spin-selective excitation and detection methods will enable the realization of compact, mechanical unidirectional components.

Here we have demonstrated scalable and reliable on-chip devices that support two-dimensional topological phenomena. These phenomena can be employed for stable and compact ultrasound and radio-frequency signal processing. With advanced nanofabrication techniques, more sophisticated structures can be realized to design other types of topological device, based on perturbative metamaterials design methods, for example[7,28]. Moreover, frequency tunability in nanoelectromechanical resonators via electrostatic forces[29,30] will be of use in electrically tunable devices[23] and actively reconfigurable topological channels[31].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0764-0.

Received: 26 June 2018; Accepted: 7 October 2018;
Published online 12 December 2018.

1. Olsson, R. H. III & El-Kady, I. Microfabricated phononic crystal devices and applications. *Meas. Sci. Technol.* **20**, 012002 (2009).
2. Baboly, M. G. et al. Demonstration of acoustic waveguiding and tight bending in phononic crystals. *Appl. Phys. Lett.* **109**, 183504 (2016).
3. Hasan, M. Z. & Kane, C. L. Topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
4. Süsstrunk, R. & Huber, S. D. Observation of phononic helical edge states in a mechanical topological insulator. *Science* **349**, 47–50 (2015).
5. Nash, L. M. et al. Topological mechanics of gyroscopic metamaterials. *Proc. Natl Acad. Sci. USA* **112**, 14495–14500 (2015).
6. Mitchell, N. P., Nash, L. M., Hexner, D., Turner, A. M. & Irvine, W. T. M. Amorphous topological insulators constructed from random point sets. *Nat. Phys.* **14**, 380–385 (2018).
7. Serra-Garcia, M. et al. Observation of a phononic quadrupole topological insulator. *Nature* **555**, 342–345 (2018).
8. Yu, S. et al. Elastic pseudospin transport for integratable topological phononic circuits. *Nat. Commun.* **9**, 3072 (2018).
9. He, C. et al. Acoustic topological insulator and robust one-way sound transport. *Nat. Phys.* **12**, 1124–1129 (2016).
10. Lu, J. et al. Observation of topologically valley transport of sound in sonic crystals. *Nat. Phys.* **13**, 369–374 (2017).
11. Yves, S., Fleury, R., Lemoult, F., Fink, M. & Lerosey, G. Topological acoustic polaritons: robust and sound manipulation at the subwavelength scale. *New J. Phys.* **19**, 075003 (2017).
12. He, H. et al. Topological negative refraction of surface acoustic waves in a Weyl phononic crystal. *Nature* **560**, 61–64 (2018).
13. Peano, V., Brendel, C., Schmidt, M. & Marquardt, F. Topological phases for sound and light. *Phys. Rev. X* **5**, 031011 (2015).
14. Brendel, C., Peano, V., Painter, O. J. & Marquardt, F. Pseudomagnetic fields for sound at the nanoscale. *Proc. Natl Acad. Sci. USA* **114**, E3390–E3395 (2017).
15. Brendel, C., Peano, V., Painter, O. J. & Marquardt, F. Snowflake phononic topological insulator at the nanoscale. *Phys. Rev. B* **97**, 020102 (2018).
16. Huang, X. M. H., Zorman, C. A., Mehregany, M. & Roukes, M. L. Nanoelectromechanical systems: nanodevice motion at microwave frequencies. *Nature* **421**, 496 (2003).
17. Craighead, H. G. Nanoelectromechanical systems. *Science* **290**, 1532–1535 (2000).
18. De Alba, R. D. et al. Tunable phonon-cavity coupling in graphene membranes. *Nat. Nanotechnol.* **11**, 741–746 (2016).
19. Faust, T., Rieger, J., Seitner, M. J., Kotthaus, J. P. & Weig, E. M. Coherent control of a classical nanomechanical two-level system. *Nat. Phys.* **9**, 485–488 (2013).
20. Okamoto, H. et al. Coherent phonon manipulation in coupled mechanical resonators. *Nat. Phys.* **9**, 485–488 (2013).
21. Hatanaka, D., Mahboob, I., Onomitsud, K. & Yamaguchi, H. Phonon waveguides for electromechanical circuits. *Nat. Nanotechnol.* **9**, 520–524 (2014).
22. Hatanaka, D., Dodel, A., Mahboob, I., Onomitsud, K. & Yamaguchi, H. Phonon propagation dynamics in band-edngineered one-dimensional phononic crystal waveguides. *New J. Phys.* **17**, 113032 (2015).
23. Cha, J. & Daraio, C. Electrical tuning of elastic wave propagation in nanomechanical lattices at MHz frequencies. *Nat. Nanotechnol.* **13**, 1016–1020 (2018).
24. Kurosu, M., Hatanaka, D., Onomitsu, K. & Yamaguchi, H. On-chip temporal focusing of elastic waves in a phononic crystal waveguide. *Nat. Commun.* **9**, 1331 (2018).
25. Wu, L. & Hu, X. Scheme for achieving a topological photonic crystal by using dielectric material. *Phys. Rev. Lett.* **114**, 223901 (2015).
26. Yves, S. et al. Crystalline metamaterials for topological properties at subwavelength scales. *Nat. Commun.* **8**, 16023 (2017).
27. Bernevig, B. A., Hughes, T. L. & Zhang, S. C. Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science* **314**, 1757–1761 (2006).
28. Matlack, K. H., Serra-Garcia, M., Palermo, A., Huber, S. D. & Daraio, C. Designing perturbative metamaterials from discrete models. *Nat. Mater.* **17**, 323–328 (2018).
29. Chen, C. et al. Graphene mechanical oscillators with tunable frequency. *Nat. Nanotechnol.* **8**, 923–927 (2013).
30. Bagci, T. et al. Optical detection of radio waves through a nanomechanical transducer. *Nature* **507**, 81–85 (2014).
31. Cheng, X. et al. Robust reconfigurable electromagnetic pathways within a photonic topological insulator. *Nat. Mater.* **15**, 542–548 (2016).

Acknowledgements J.C. and C.D. acknowledge partial support for this project from NSF EFRI award number 1741565, and the Kavli Nanoscience Institute at Caltech. K.W.K. acknowledges support for this project from the "Overseas Research Program for Young Scientists" programme through the Korea Institute for Advanced Study (KIAS).

Reviewer information *Nature* thanks S. Iwamoto and the other anonymous reviewers for their contribution to the peer review of this work.

Author contributions J.C. and C.D. conceived the idea of the research. J.C. designed and fabricated the samples, built the experimental setups and performed the measurements. J.C. also performed all numerical simulations. J.C. and K.W.K. performed the theoretical studies. J.C., K.W.K. and C.D. analysed the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0764-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0764-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Sample fabrication. The fabrication process begins with a pattern transfer by electron beam lithography and development of a PMMA resist in a MIBK:IPA = 1:3 solution. The excitation electrodes, made of a Au (45 nm)/Cr (5 nm) layer, are deposited on a 100-nm-LPCVD (low-pressure chemical vapour deposition) silicon nitride (SiN_x)/140-nm thermal SiO_2/525-μm highly doped Si wafer, followed by a lift-off process in acetone. A second electron beam lithography step, with ZEP 520A electron-beam resist, is then performed to create the pattern of etch holes (with diameter 500 nm) arranged in the extended honeycomb lattices (Fig. 1a and Extended Data Fig. 1). We use an ICP-reactive ion etch, to drill the holes on the SiN_x layer. After we finish the etching of the holes, we immerse the samples in a buffered oxide etchant solution for about 45–46 min, to partially etch the thermal SiO_2 underneath the SiN_x device layer. The etching duration determines the diameter of the etching circles, r (Extended Data Fig. 1). Detailed fabrication methods can be found in ref. [23]. Different samples were fabricated for the measurements in Figs. 2 and 3, with slightly different etching times. These differences in fabrication lead to a small change in the operating frequencies.

Experiments. The flexural motions of the membranes are measured using a home-built optical interferometer (HeNe laser, wavelength 633 nm) with a balanced homodyne method. The measurements are performed at room temperature and a vacuum pressure of $p < 10^{-6}$ mbar. The optical path length difference between the reference and the sample arms is stabilized by actuating a reference mirror. This mirror is mounted on a piezoelectric actuator that is controlled by a proportional integral derivative (PID) controller. The motion of the membranes is electrostatically excited by simultaneously applying a constant and a time-varying voltage through a bias tee (Mini-circuits, ZFBT-6GW+). The intensity of the interfered light from the reference mirror and the sample is measured using a balanced photodetector, which is connected to a high-frequency lock-in amplifier (Zurich instruments, UHFLI). The measurement position, monitored via a complementary metal–oxide–semiconductor (CMOS) camera, can be controlled by moving a vacuum chamber mounted on a motorized XY linear stage.

For the dispersion curve measurements in Fig. 2 and in the Extended Data Fig. 5, we measure (at steady-state) frequency responses of 10–20 MHz of 20 scanned sites along the edge waveguide. The scanning step is the one-dimensional lattice spacing, $a = 18\,\mu m$. The lock-in amplifier (Zurich instruments, UHFLI) allows us to measure the amplitude responses and the phase differences between the measured signal and the excitation source. To plot the frequency dispersion, we perform fast Fourier transformation of the amplitude × sin(phase) data. The amplitude-only data and the phase-considered data are shown in Extended Data Fig. 3a and b.

For transient measurements in Figs. 3 and 4, we send a chirped signal (AWG module in UHFLI) and measure the signal with an oscilloscope (Tektronix, DPO3034). As the signal is invisible for a low-excitation amplitude, we first filter the radio-frequency output signals from the photodetector with a passive band-pass filter (6–22 MHz bandwidth) and an average of 512 datasets in the time-domain. For robustness measurements (Fig. 3), we send a pulse containing frequency content ranging from 12.8 MHz to 15.8 MHz, by applying $V_{DC} = 15$ V and $V_P = 30$ mV to the excitation electrode. We then perform post-signal processing to extract signals of interest, by applying a Burtterworth filter for different centre frequencies with 0.3-MHz bandwidth. For pseudospin-dependent transport measurements, we use a pulse (14–16 MHz) and applied $V_{DC} = 15$ V and $V_P = 22.5$ mV. We apply a Burtterworth filter with 15.2-MHz centre frequency and 0.5-MHz bandwidth.

Numerical simulations. We perform finite-element simulations to calculate the phononic frequency dispersion curves using COMSOL multiphysics. We employ the pre-stressed eigenfrequency analysis module in membrane mechanics. We also consider geometric nonlinearity, to reflect the effects of residual stresses. The physical properties of SiN_x used in the simulations are density 3,000 kg m^{-3}, Young's modulus 290 GPa, Poisson ratio 0.27 and isotropic in-place residual stress 50 MPa. The lattice parameter, a, is chosen to be 18 μm. We calculate frequency dispersion curves for various unit cell geometries with different w ranging from 5.5 μm to 6.5 μm. The centre hexagon and the six corners of each unit cell are fixed, owing to the presence of unetched SiO_2 (light-grey regions in the SEM images in Fig. 2b–d). The radii of the etched circles are set to $r = 4.9\,\mu m$ (Fig. 1 and Extended Data Fig. 2). We apply Bloch periodic conditions to the six sides of a unit cell, $u(\boldsymbol{r}+\boldsymbol{R}) = u(\boldsymbol{r})\exp(i\boldsymbol{q}\cdot\boldsymbol{R})$, via Floquet periodicity in COMSOL. Here, $u(\boldsymbol{r})$ is a periodic displacement function, $\boldsymbol{r}$ is a position within a unit cell, $\boldsymbol{R}$ is a lattice translation vector, and $\boldsymbol{q}$ is a wavevector. We calculate the dispersion curves along the boundary of the irreducible Brillouin zone ΔMΓK in Fig. 1c.

We also numerically calculate the frequency dispersion curves of the edge states to validate the topological behaviours. As we are interested in one-dimensional dispersion along the interface, we build a strip-like super cell with 18-μm periodicity. Each topological phase ($w = 6.0 \pm 0.5\,\mu m$) spans about ±160 μm from the interface in the direction perpendicular to the interface. We calculate the frequency dispersion by applying one-dimensional Bloch periodic conditions.

***k·p* perturbation method and Bernevig–Hughes–Zhang model.** The equation of motion for a plate of thickness h is

$$D\nabla^4 W = -\rho h\frac{\partial^2 W}{\partial t^2}$$

Here, $D = Eh^3/[12(1-\nu^2)]$ is the bending stiffness, ρ is density, h is the plate thickness and W is the plate displacement in the z direction. By inserting a Bloch function $W_{n,\boldsymbol{q}}(\boldsymbol{r}) = e^{i(\boldsymbol{q}\cdot\boldsymbol{r}-\omega t)}Y_{n,\boldsymbol{q}}(\boldsymbol{r})$ to the plate equation, we obtain $[D/(\rho h)]HY_{n,\boldsymbol{q}} = \omega_{n,\boldsymbol{q}}^2 Y_{n,\boldsymbol{q}}$. Here, $\boldsymbol{r}$ is a position within a unit cell, $\boldsymbol{q}$ is a wavevector, n is a band index, $\omega_{n,\boldsymbol{q}}$ is an eigenfrequency and $Y_{n,\boldsymbol{q}}(\boldsymbol{r})$ is a periodic displacement function. The operator H is given by

$$\begin{aligned} H=&\frac{\partial^4}{\partial x^4}+2\frac{\partial^4}{\partial x^2\partial y^2}+\frac{\partial^4}{\partial y^4}\\ &+|\boldsymbol{q}|^4+4i\boldsymbol{q}\cdot\boldsymbol{\nabla}(\nabla^2)-4i|\boldsymbol{q}|^2\boldsymbol{q}\cdot\boldsymbol{\nabla}-2|\boldsymbol{q}|^2\nabla^2-4\boldsymbol{q}\cdot\boldsymbol{\nabla}(\boldsymbol{q}\cdot\boldsymbol{\nabla})\\ &=H^0+H'(\boldsymbol{q}) \end{aligned}$$

The equation of motion that describes a plate vibration includes a square of the Laplacian operator ∇^4, so the higher-order wavevector terms arise, accordingly. Here, we consider the $\boldsymbol{q}$-dependent terms, $H'(\boldsymbol{q})$, as a small perturbation, as we are only interested in the behaviours near the Γ point ($\boldsymbol{q} = \boldsymbol{0}$) of the Brillouin zone. This perturbation term is equivalent to the ***k·p*** term in the ***k·p*** perturbation theory in quantum mechanics.

To obtain a Hamiltonian matrix on the pseudospin subspace, expressed in the basis $\{Y_{p_\pm,0}, Y_{d_+,0}, Y_{p_-,0}, Y_{d_-,0}\}$. We define the pseudospin states as $Y_{p_\pm,0} = (Y_{p_x,0} \pm iY_{p_y,0})/\sqrt{2}$ and $Y_{d_\pm,0} = (Y_{d_{x^2-y^2},0} \pm iY_{d_{xy},0})/\sqrt{2}$. The Bloch states, $Y_{p_x,0}$, $Y_{p_y,0}$, $Y_{d_{x^2-y^2},0}$ and $Y_{d_{xy},0}$, are the eigenstates of the operator at the Γ point. We neglect the cubic and quartic wavevector terms. The matrix elements can be calculated from

$$\begin{aligned} H_{n,m} =& \oint_{\text{unit cell}} Y_{n,0}^* H^0 Y_{m,0}\mathrm{d}A\\ &+\oint_{\text{unit cell}} Y_{n,0}^*[4i\boldsymbol{q}\cdot\boldsymbol{\nabla}(\nabla^2)-2|\boldsymbol{q}|^2\nabla^2-4\boldsymbol{q}\cdot\boldsymbol{\nabla}(\boldsymbol{q}\cdot\boldsymbol{\nabla})]Y_{m,0}\mathrm{d}A \end{aligned}$$

where the indices n and m denote the pseudospin states. Neglecting the off-block-diagonal elements as they contribute as a higher-order perturbation, the matrix is expressed as

$$H(\boldsymbol{q}) = \alpha(\boldsymbol{q})I + \begin{bmatrix} \Omega(\boldsymbol{q}) & 0 \\ 0 & \Omega^*(-\boldsymbol{q}) \end{bmatrix}$$

where

$$\alpha(\boldsymbol{q}) = \frac{\omega_p^2+\omega_d^2}{2} + \left(\frac{A+B}{2}\right)(q_x^2+q_y^2)$$

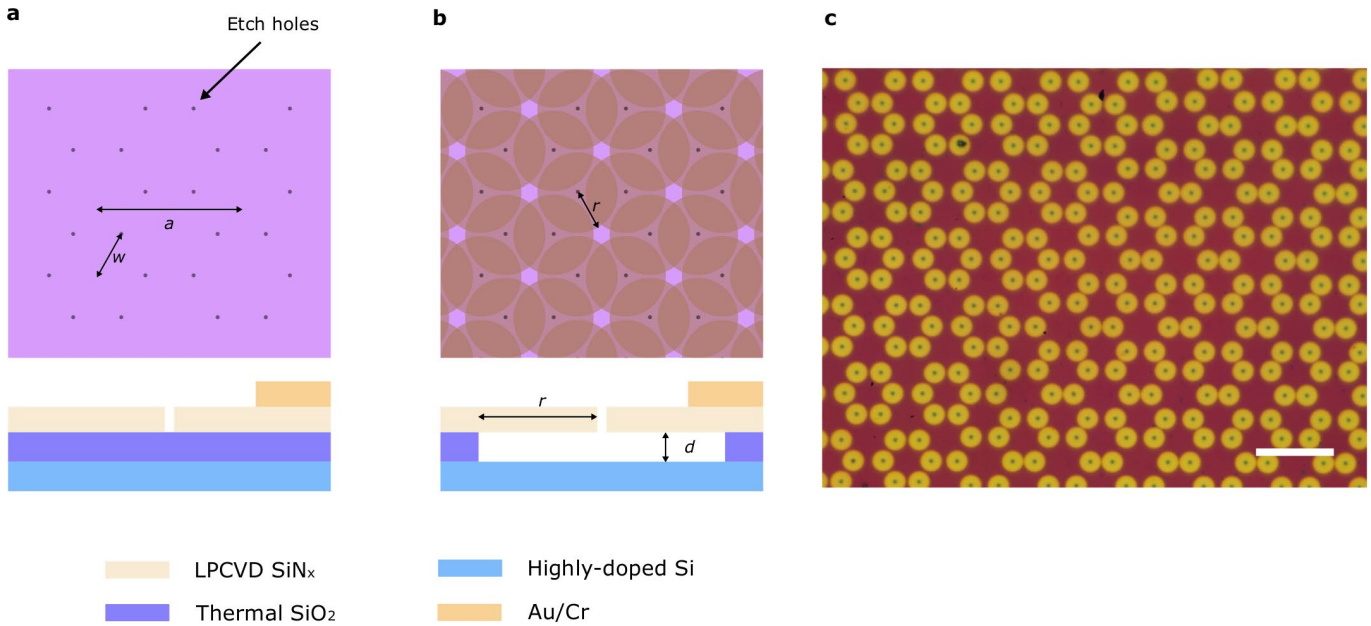
$$\Omega(\boldsymbol{q}) = \begin{bmatrix} \dfrac{\omega_p^2-\omega_d^2}{2}+\left(\dfrac{A-B}{2}\right)(q_x^2+q_y^2) & C(q_x+iq_y) \\ C^*(q_x-iq_y) & -\left(\dfrac{\omega_p^2-\omega_d^2}{2}\right)-\left(\dfrac{A-B}{2}\right)(q_x^2+q_y^2) \end{bmatrix}$$

Here, the asterisk denotes the complex conjugate. The eigenvalues of H^0 are $\omega_p^2 = \omega_{p_x}^2 = \omega_{p_y}^2$ and $\omega_d^2 = \omega_{d_{x^2-y^2}}^2 = \omega_{d_{xy}}^2$. This matrix shows a similar form to the Hamiltonian matrix from Bernevig–Hughes–Zhang model[27] that describes the quantum spin Hall effect in two-dimensional systems. This result confirms that the system can support two different pseudospins.

Data availability

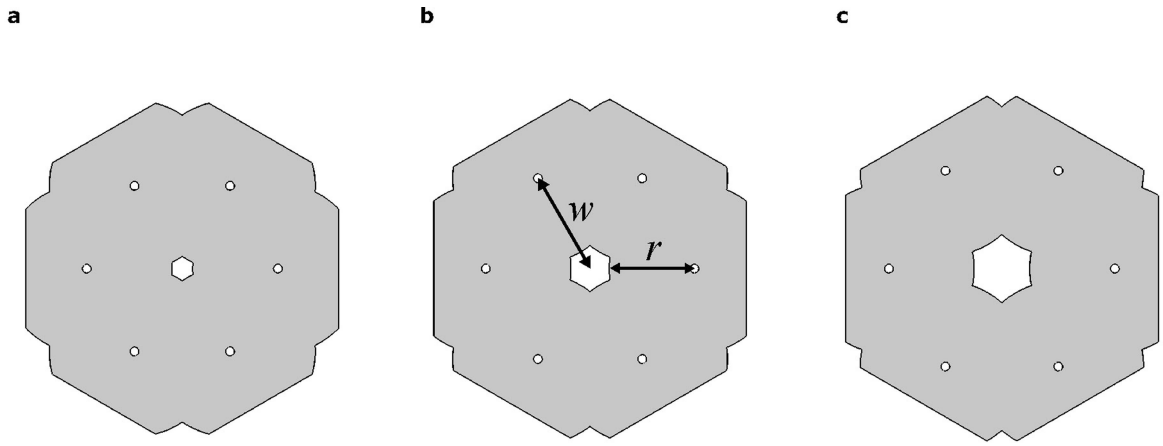
The data that support the findings of this study are available from the corresponding author upon reasonable request.

© 2018 Springer Nature Limited. All rights reserved.



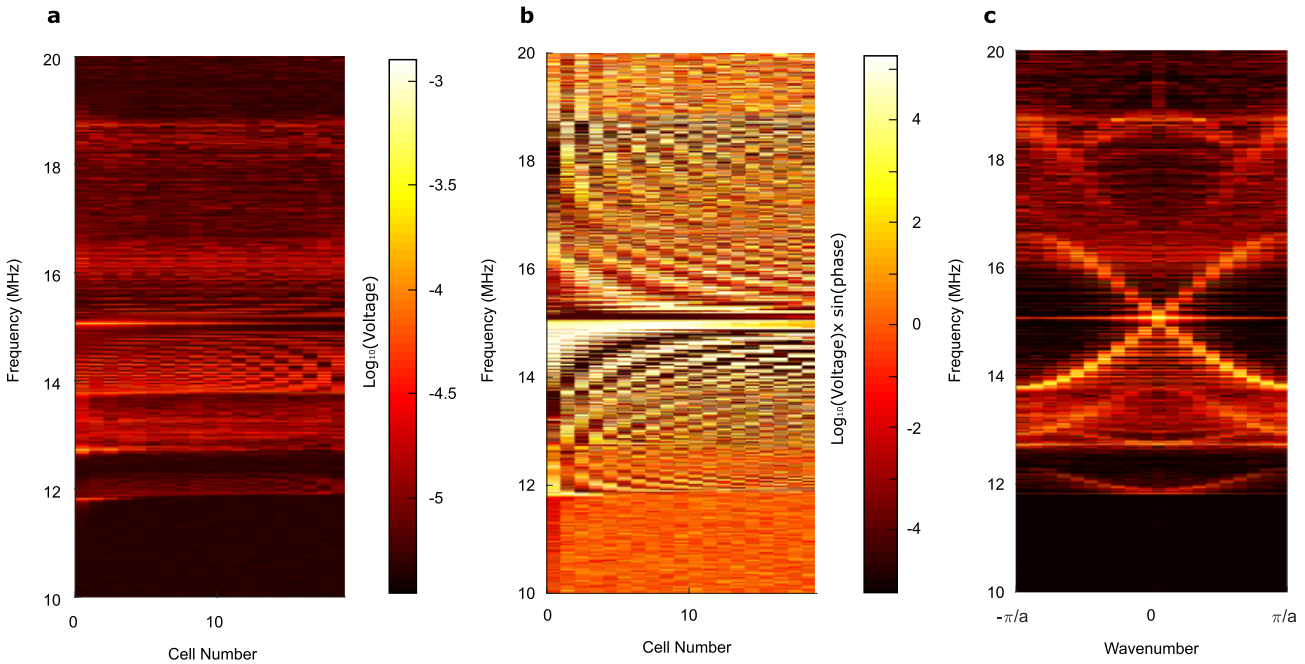
Extended Data Fig. 1 | Fabrication. a, Etch holes are periodically arranged in an extended honeycomb lattice. *a* is a lattice parameter and *w* is the distance between two neighbouring etch holes. The bottom picture shows the cross-sectional view. **b**, After a sample is immersed in a buffered oxide etchant, the thermal SiO_2 is radially etched from the etch holes. The etching paths are illustrated by the brown circles with radius *r*. The overlap between two membranes affects the coupling strength, by controlling the distance *w*. **c**, Optical microscope image of a partially etched samples. The yellow circles represent free-standing SiN_x membranes and the darker/purple region is SiN_x/SiO_2. Scale bar, 18 μm.

© 2018 Springer Nature Limited. All rights reserved.



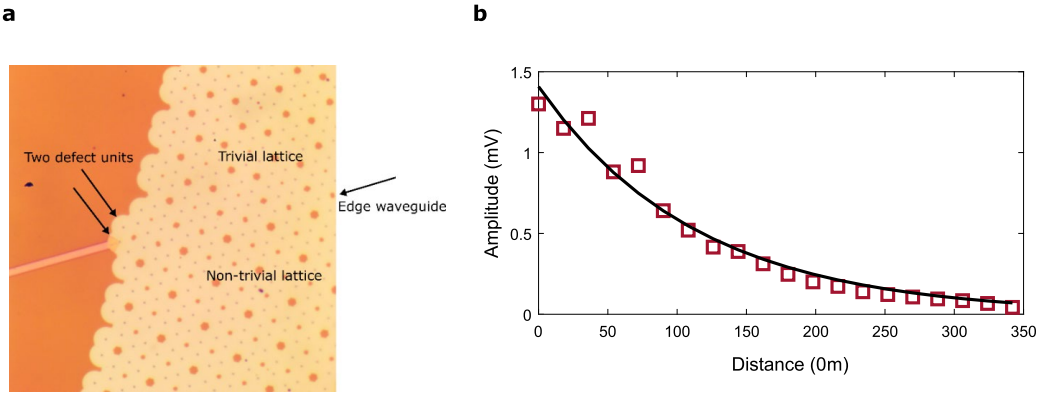
Extended Data Fig. 2 | Unit cell structures for the different topological phases. a–c, Unit cell geometries for $w = 5.5\,\mu m$ (**a**), $w = 6.0\,\mu m$ (**b**) and $w = 6.5\,\mu m$ (**c**). Here, $r = 4.9\,\mu m$ is the etching distance from the centre of the etch holes. The central hexagon and the six corners represent the regions of unetched thermal SiO_2. These are modelled as fixed boundaries in finite element simulations.

© 2018 Springer Nature Limited. All rights reserved.



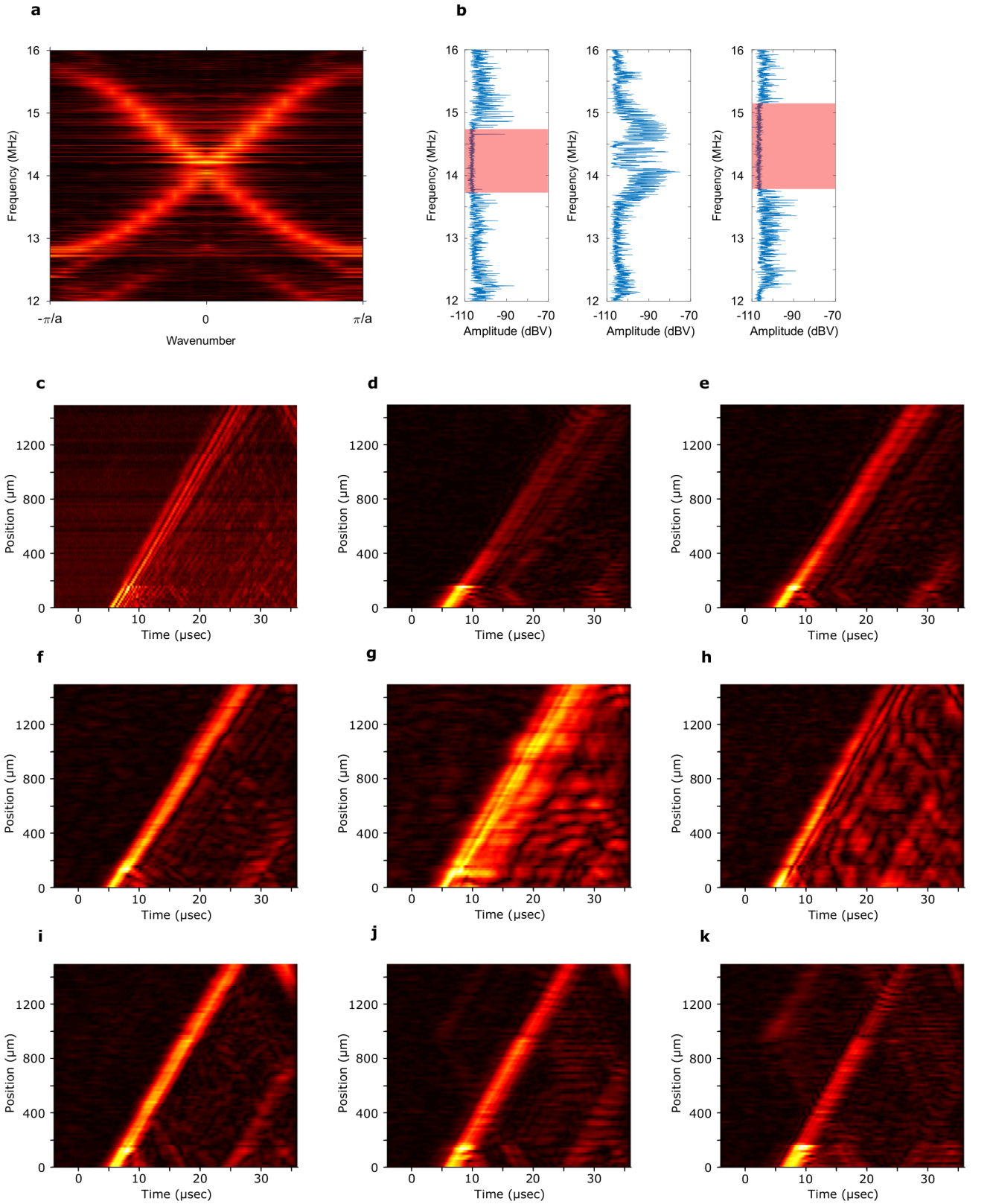
Extended Data Fig. 3 | Experimental characterization of the straight topological edge waveguide. a, Frequency responses (amplitude only) obtained from a spatial scan of 20 sites of the straight edge waveguide. The scanning period is 18 μm. The intensity decay around 15 MHz represents a defect mode. **b**, Amplitude × sin(phase) responses obtained from the spatial scan. **c**, Frequency dispersion obtained from Fourier transformation of the data shown in **b**.

© 2018 Springer Nature Limited. All rights reserved.



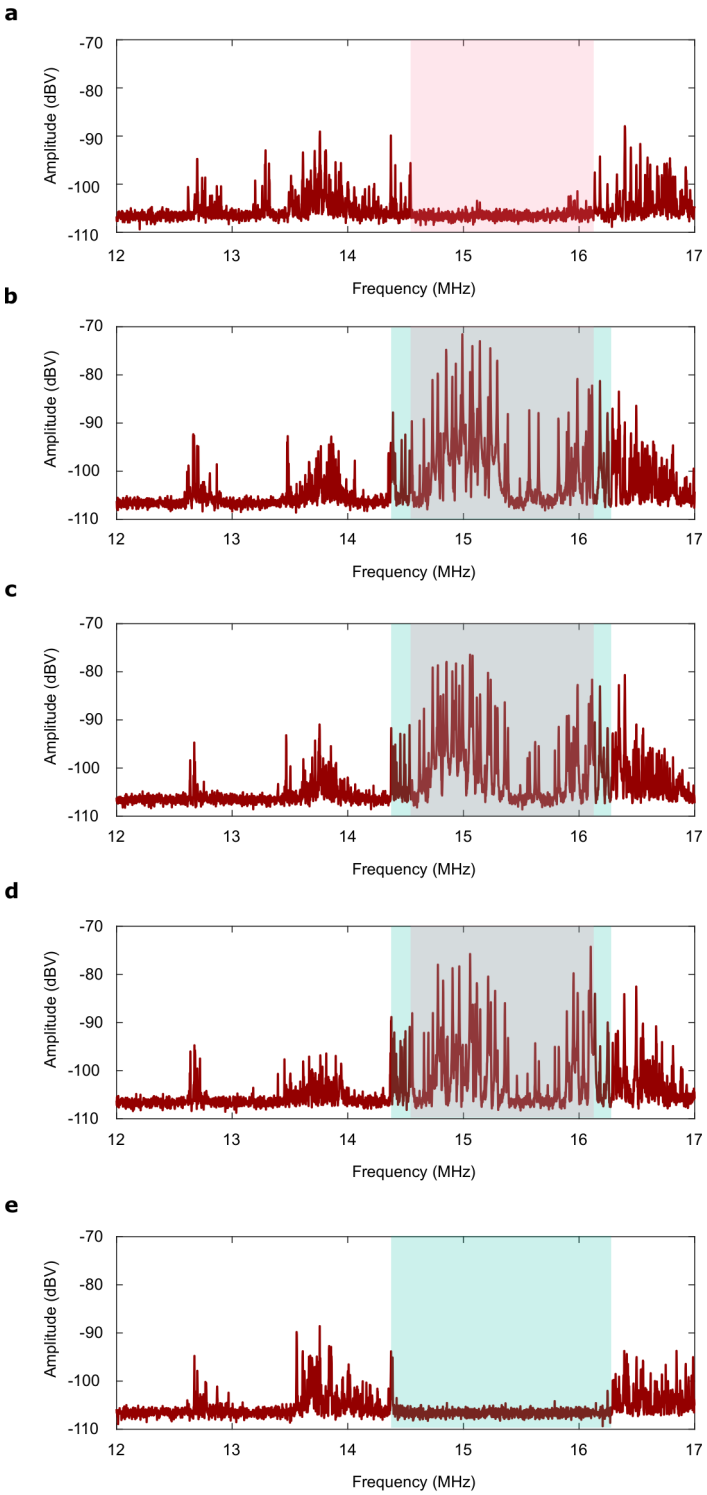
Extended Data Fig. 4 | Characteristics of the defect mode. a, Optical microscope image of the excitation region and the defective unit including the gold electrode. **b**, Experimental data for the amplitude decay of the defect mode. The red squares denote the experimental data and the black solid line represents a fitting function. Here, the fitting parameters a and b are 1.408 mV and 114.1292 μm.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Characteristics of the distorted waveguide. **a**, Frequency dispersion curve along the edge waveguide. **b**, Frequency responses of the bulk of the non-trivial phase (left), the output of the edge waveguide (middle), and the bulk of the trivial phase (right). **c**, A spatiotemporal response of unfiltered propagating pulses with broadband frequencies ranging from 12.8 MHz to 15.8 MHz. **d**–**k**, Filtered spatiotemporal responses of the propagating pulses with different centre frequencies. The bandwidth of the pulses is 0.3 MHz. The centre frequencies are 13.48 MHz (**d**), 13.68 MHz (**e**), 13.85 MHz (**f**), 14.1 MHz (**g**), 14.35 MHz (**h**), 14.75 MHz (**i**), 14. 975 MHz (**j**) and 15.23 MHz (**k**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Frequency responses of the NEMM with the pseudospin filter configuration. a–e, Frequency responses of the bulk of the non-trivial phase ($w = 6.5\,\mu m$; **a**), output 1 (**b**), output 2 (**c**), output 3 (**d**) and the bulk of the trivial phase ($w = 5.5\,\mu m$; **e**), shown in Fig. 4a. The light-red and light-blue regions represent the bandgaps of the non-trivial and trivial phases, respectively.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0754-2

Reversible superdense ordering of lithium between two graphene sheets

Matthias Kühne[1,6,8], Felix Börrnert[2,8], Sven Fecher[1], Mahdi Ghorbani-Asl[3], Johannes Biskupek[2], Dominik Samuelis[1,7], Arkady V. Krasheninnikov[3,4,5], Ute Kaiser[2]* & Jurgen H. Smet[1]*

Many carbon allotropes can act as host materials for reversible lithium uptake[1,2], thereby laying the foundations for existing and future electrochemical energy storage. However, insight into how lithium is arranged within these hosts is difficult to obtain from a working system. For example, the use of in situ transmission electron microscopy[3–5] to probe light elements (especially lithium)[6,7] is severely hampered by their low scattering cross-section for impinging electrons and their susceptibility to knock-on damage[8]. Here we study the reversible intercalation of lithium into bilayer graphene by in situ low-voltage transmission electron microscopy, using both spherical and chromatic aberration correction[9] to enhance contrast and resolution to the required levels. The microscopy is supported by electron energy-loss spectroscopy and density functional theory calculations. On their remote insertion from an electrochemical cell covering one end of the long but narrow bilayer, we observe lithium atoms to assume multi-layered close-packed order between the two carbon sheets. The lithium storage capacity associated with this superdense phase far exceeds that expected from formation of LiC_6, which is the densest configuration known under normal conditions for lithium intercalation within bulk graphitic carbon[10]. Our findings thus point to the possible existence of distinct storage arrangements of ions in two-dimensional layered materials as compared to their bulk parent compounds.

Figure 1a shows a schematic of our devices, all of which are supported by Si_3N_4-covered Si substrates; the Si_3N_4 forms a 40 μm × 40 μm membrane at the centre of the Si chip. A bilayer graphene flake is exfoliated from natural graphite and etched into a Hall bar shape. One end of the flake is connected to a counter electrode on the Si_3N_4 surface via a Li-ion conducting solid polymer electrolyte that has been encapsulated in a thin layer of SiO_x to avoid outgassing and oxidation in air. This setup with an electrochemical cell at the end of the flake allows for the controlled reduction/oxidation of bilayer graphene according to[1]:

$$x\mathrm{Li}^+ + xe^- + \mathrm{C}_n \rightleftharpoons \mathrm{Li}_x\mathrm{C}_n \quad (1)$$

Using a procedure similar to that in refs [11,12], we trigger lithiation (delithiation) by applying a positive voltage $U_G = 5$ V ($U_G = 0$ V) to the counter electrode with respect to the bilayer graphene. A grounded current lead to the latter serves as a source/sink for the electrons required to facilitate the reversible intercalation of Li ions at the electrolyte-covered end of the bilayer graphene bar. Intercalated Li exhibits rapid lateral diffusion that tends to establish and maintain an even distribution of Li throughout the bilayer[13]. Hence, one may study its ordering in a region well separated from the electrolyte by in situ TEM (Fig. 1a, b), thereby also preventing exposure of the electrolyte to the electron beam. In the region probed by TEM, bilayer graphene is suspended over a hole in the Si_3N_4 membrane. Metallic contacts to the bilayer allow monitoring of its resistivity $\rho_{xx} = U_{xx}/I$ (where U_{xx} is the longitudinal voltage drop and I is the applied current) in the electrolyte-uncovered region (Fig. 1a). During subsequent lithiation/delithiation cycles, we typically observe reversible changes in ρ_{xx} (Fig. 1c). These relate to changes in the local Li concentration via finite electronic charge transfer[12,13]. The decrease of ρ_{xx} during lithiation reflects an increase in electron density, characteristic of ambipolar diffusion of electron–ion pairs into the probed area[13]. During delithiation, Li ions and electrons leave the bilayer, thereby restoring its initial resistivity value. The exact time evolution of ρ_{xx} depends on the kinetics of several (uncontrolled) processes, related to, for example, ionic transport within the electrolyte and across the solid electrolyte interphase. Yet the behaviour shown in Fig. 1c is a qualitative characteristic of reversible Li intercalation in bilayer graphene[13]. In the following, we present in situ TEM data obtained in the unique spherical- and chromatic-aberration corrected SALVE (Sub-Ångström Low-Voltage Electron microscopy) instrument[9]. We work at an electron acceleration voltage of 80 kV, just below the threshold for knock-on damage of C atoms in graphene[14]. Under these conditions, the instrument delivers sub-ångström resolution in the images. Further details can be found in Methods (and also in Extended Data Figs. 1, 2).

Figure 2a shows a TEM image of pristine bilayer graphene. The inset depicts its Fourier transform and confirms the known value of the in-plane lattice constant $a_C = 2.46$ Å. The image marks the beginning of a series acquired during lithiation of a bilayer graphene device (Fig. 2a–c and Supplementary Video 1), but it is representative of the sample state before application of the bias voltage $U_G = 5$ V. In Fig. 2b, acquired after 170 s, a second crystal lattice has appeared in the lower half of the probed area. White dashed lines demarcate its boundary on the left and right of the image. The image in Fig. 2c is recorded at $t = 288$ s. The additional crystal structure now extends throughout almost the whole field of view. Figure 2d is the Fourier transform of Fig. 2b. When we compare Fig. 2d with the Fourier transform of pristine bilayer graphene (Fig. 2a inset), we identify three sets of additional signals, highlighted in red, green and blue. These attest to hexagonal crystalline order (as for graphene), but with an in-plane lattice constant of 3.1 Å. In Fig. 2e, the spatial distribution of these additional signals is mapped. This allows three grains to be discerned, none of which are aligned with the encapsulating graphene lattice. In Fig. 2f the same Fourier transform is shown but with a von Hann filter applied to minimize the streaks. The highlighted signals stem from both bilayer graphene (cyan) and the additional crystalline phase (green) as well as moiré artefacts (magenta) and their origin (bold arrows). Figure 2g is a Fourier-filtered (Methods and Extended Data Fig. 3) version of Fig. 2b, where the graphene lattice, as well as the moiré effects, have been removed (see Fig. 2h, i, respectively, for a magnified view before and after filtering). These images offer a direct view of the encapsulated crystal. The observed contrast in the images also suggests regions of different thickness even within a single grain. This is worked out in detail in Methods and Extended Data Fig. 4.

[1]Max Planck Institute for Solid State Research, Stuttgart, Germany. [2]Materialwissenschaftliche Elektronenmikroskopie, Universität Ulm, Ulm, Germany. [3]Institute of Ion Beam Physics and Materials Research, Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany. [4]Department of Applied Physics, Aalto University, Aalto, Finland. [5]National University of Science and Technology MISiS, Moscow, Russia. [6]Present address: Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Present address: Heraeus Battery Technology, Hanau, Germany. [8]These authors contributed equally: M. Kühne, F. Börrnert. *e-mail: ute.kaiser@uni-ulm.de; j.smet@fkf.mpg.de

© 2018 Springer Nature Limited. All rights reserved.

To narrow down the chemical composition of the additional crystal structure, electron energy-loss spectroscopy (EELS) data has been acquired on bilayer graphene before and during lithiation (Fig. 2j). Before lithiation ('pristine'), we exclusively observe the C K-edge at 284 eV. During lithiation, the Li K-edge at 55 eV is additionally detected from regions characterized by a Fourier transform like that shown in Fig. 2d. On the basis of the absence of other signals in the explored energy range of 0–800 eV—namely, Si ($L_{2,3}$-edge at 99 eV), S ($L_{2,3}$-edge at 165 eV), N (K-edge at 400 eV), O (K-edge at 532 eV and F (K-edge at 685 eV)—we discarded those elements as principal constituents of the new crystalline phase. Likewise, Ti and Pt (electrode material) can be ruled out in view of the observed light atomic contrast in Fig. 2 for the case of Pt and the absence of the distinct Ti $L_{2,3}$-edge at 456 eV in the EELS data. Although it is inherently impossible to exclude H or C, the crystalline phase formed during lithiation is likely to consist of pure Li. The low onset in the energy of the Li K-edge supports this assertion[15,16]. Also, a Li plasmon mode[17,18] appears near 9 eV (Extended Data Fig. 5a). Though the shape of the Li-K edge[15,16] resembles that of Li_2O/LiOH, stoichiometric Li_2O/LiOH can be disqualified, since we would then expect both a more pronounced O-K edge and an imaging contrast comparable in strength to the encapsulating graphene[19]. Good agreement with experiment is attained when calculating the Li-K edge shape of graphene-encapsulated Li multilayers (Methods and Extended Data Fig. 5b). We do not rule out the presence of trace oxygen, also suggested by the occasional observation of a weak shoulder near 30 eV (Extended Data Fig. 5a), previously attributed to oxidized lithium[17,18]. Yet, the extracted in-plane lattice constant of 3.1 Å matches that[20] of close-packed Li. This coincidence is surprising, since normally very low temperatures and/or extreme pressures are required for Li to assume this superdense ordering[21].

To test whether formation of a dense, multi-layered Li crystal in bilayer graphene is conceivable, first-principles calculations were carried out (Methods). The chemical potential of Li atoms in bulk close-packed phases was evaluated. All energies below are given with respect to the hexagonal close-packed (h.c.p.) phase, which had the lowest energy in the calculations. Several layers of Li atoms inside bilayer graphene are considered (Fig. 3 and Extended Data Figs. 6–8). Because the in situ TEM studies do not reveal evidence for a change in registry of the graphene sheets (Extended Data Fig. 9), AB stacking is assumed as in the pristine device. Our main conclusions, however, hold irrespective of the stacking order. The relaxation of atomic coordinates of a single layer of Li atoms yields an energetically favoured C_6LiC_6 configuration, with Li arranged in a commensurate $(\sqrt{3} \times \sqrt{3})R30^\circ$ superstructure with a lattice constant of $a_{LiC6} = 4.26$ Å (Fig. 3a, b). Except for the graphene registry, this finding is similar to the bulk LiC_6 phase[10] that forms in graphite and the AA-stacked C_6LiC_6 phase[22]. The situation changes for a larger number of Li layers. Finite clusters and infinite (periodic) h.c.p. structures with different orientation with respect to the graphene lattice were considered (Fig. 3c–f, Extended Data Figs. 7, 8). The energies of these systems are very close to that of the C_6LiC_6 configuration (higher by only 0.01–0.05 eV per Li atom). The in-plane lattice constant a_{Li} of the Li h.c.p. bilayer and trilayer (average distances, as the positions of Li atoms are affected by nearby C atoms) is in the range 3.05–3.15 Å, matching the experimental value of 3.1 Å as well as the identical literature value[20]. These results suggest that the formation of a multilayer close-packed Li phase between graphene sheets is conceivable. Its precise stacking order may, however, differ from h.c.p. as other configurations are energetically similar (Extended Data Fig. 8d, e). These are hard to distinguish in TEM experiments, because we usually only observe the projection along one crystal direction. Moreover, bulk diffraction selection rules do not hold in the given case of an atomically thin specimen. TEM image simulations reveal comparable contrast and diffraction patterns for the two extreme cases: cubic close-packed (c.c.p.) and h.c.p. phases (Extended Data Fig. 2f, g).

By analysing the electronic structure and charge distribution, the charge transfer between Li and graphene can be estimated. Figure 3b, d, f

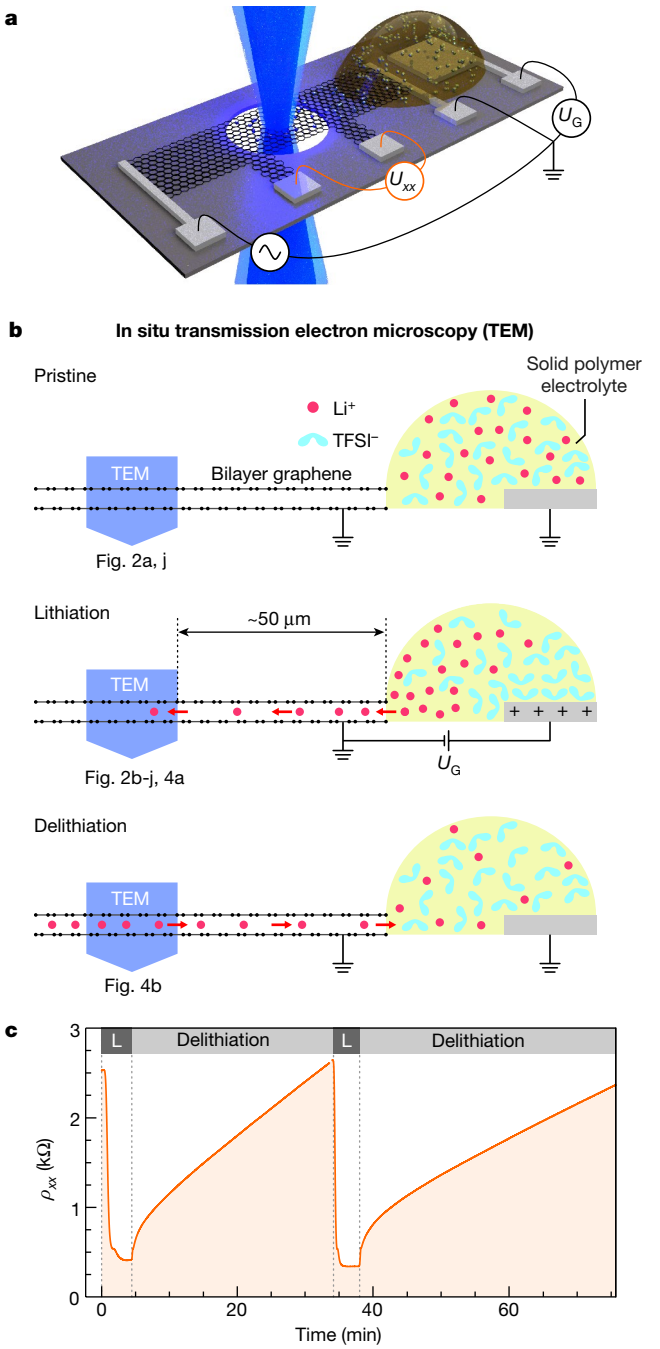


Fig. 1 | Device layout and working principle. a, Schematic of the device (not to scale). Bilayer graphene (black) on a Si_3N_4-covered Si substrate (dark grey) with a membrane at the centre of the chip is contacted by several metallic electrodes (light grey). On the right, a Li-ion (white spheres) conducting electrolyte (yellow) connects the bilayer to a metallic counter electrode to form an electrochemical cell. About 50 µm away from its electrolyte-covered end, bilayer graphene is partially suspended over a hole in the Si_3N_4, supported by the Si substrate, allowing transmission electron microscopy (TEM) investigations (electron beam illustrated in blue). **b**, Schematic side view of the device during in situ TEM: top panel, the pristine device; middle and bottom panels, the device during lithiation ($U_G = 5$ V at the counter electrode) and delithiation ($U_G = 0$ V), respectively. Ionic components of the solid polymer electrolyte are indicated by red dots and cyan shapes; see Methods for details. Reference to TEM data acquired at the respective state is given. **c**, Bilayer graphene's resistivity ρ_{xx} measured in situ during two lithiation (L)/delithiation cycles inside the SALVE microscope with the electron beam blanked. As schematically shown in **a**, a four-point probe configuration is used to imprint a small a.c. current (~) and measure the local voltage drop U_{xx} in the electrolyte-uncovered region of the device.

© 2018 Springer Nature Limited. All rights reserved.

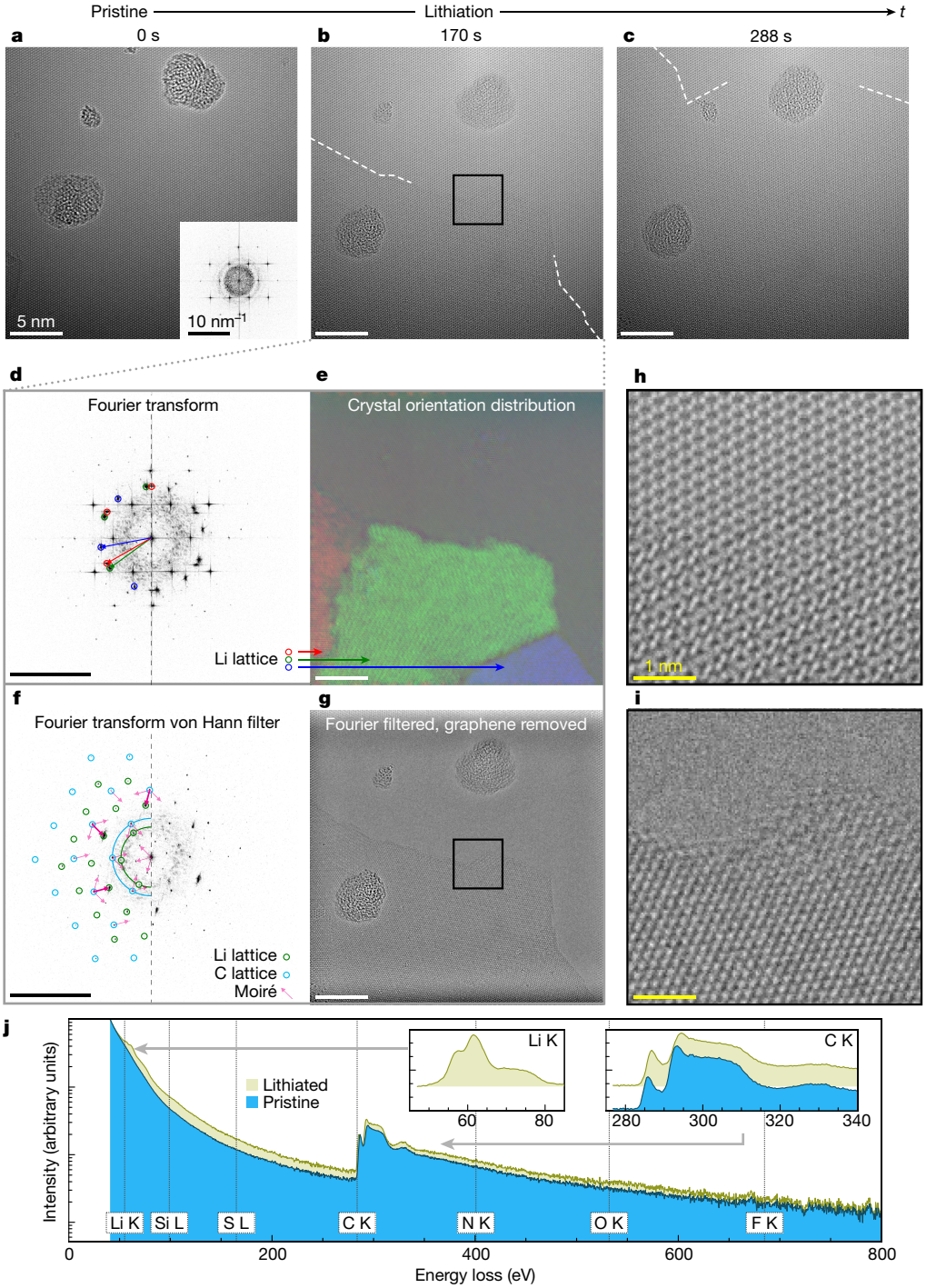


Fig. 2 | In situ TEM measurements. a–c, TEM images showing the propagating front (dashed white line) of a Li crystal forming inside bilayer graphene during lithiation. The images are acquired on the same sample area at consecutive times, as indicated at the top. Amorphous hydrocarbon adsorbates appear as blobs a few nanometres wide, located above or below the bilayer. The inset in **a** is its Fourier transform. Panels **d–g** give further information about panel **b**. **d**, Fourier transform of **b**. Three sets of spots marked in red, green and blue are rotated relative to each other as they stem from three different Li crystal grains. The cross-shaped streaks are edge artefacts from the Fourier transform. **e**, Spatial distribution of the Li grains in **b** using the colour coding from **d**. **f**, von Hann-filtered Fourier transform of **b**. Fourier transforms are point-symmetric, therefore, marks on the left side do not mask information. Signals from bilayer graphene (Li) are highlighted in cyan (green); the origins of moiré artefacts are highlighted in bold magenta. The anisotropic smearing of the Li signals is due to the sharp propagation front. Signals originating from the grains coloured red and blue in **e** are damped away by the applied von Hann filter. Half circles represent the fundamental periodicities of 0.213 nm for graphene (cyan) and 0.276 nm for Li (green). **g**, Fourier-filtered version of **b**, where the graphene lattice, as well as the moiré effects, are filtered out and only the Li crystal structure is left. The contrast at the edge of the figure is an artefact from the Fourier filter. **h**, **i**, Magnified detail from the boxed areas in **b** and **g**, respectively, showing the Li crystal edge. Scale bars of equal size are coloured identically but labelled only once in **a–i**. **j**, EELS data with logarithmic intensity scale before (blue) and during (yellow) lithiation, acquired on an area as in **a** and **c**, respectively. Highlighted are the energies of the relevant major edges. Insets show the near-edge structure of the Li and C K-edges on a linear intensity scale and after individual subtraction of an inverse-power-law background. Note that EELS data become noisy at high energies because of the exponential decay of the signal.

© 2018 Springer Nature Limited. All rights reserved.

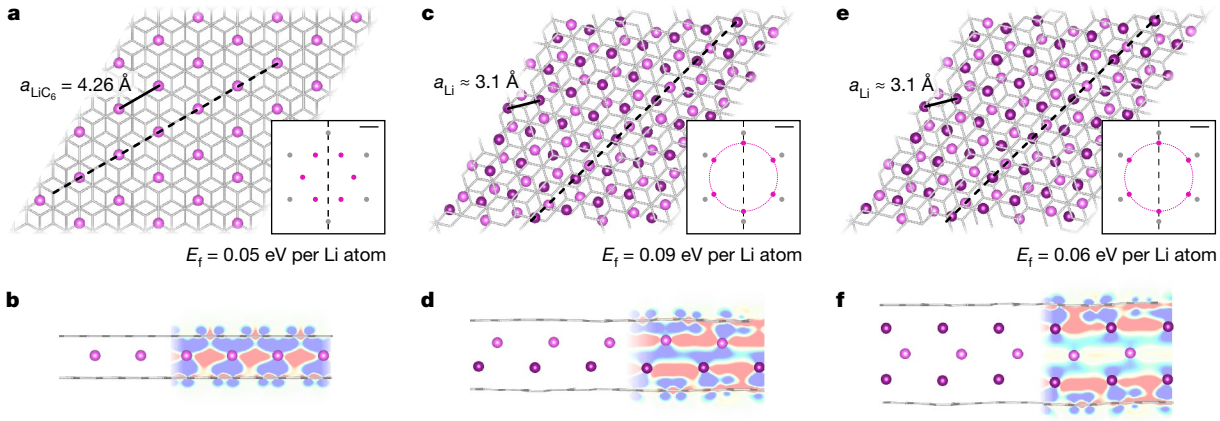


Fig. 3 | Atomistic models of Li crystals between AB-stacked graphene sheets obtained from DFT calculations. a, **b**, The 'conventional' C_6LiC_6 configuration with Li arranged in a commensurate $(\sqrt{3} \times \sqrt{3})R30^\circ$ superstructure between graphene sheets. **c**, **d**, Fully optimized bilayer Li crystal. **e**, **f**, Fully optimized trilayer Li crystal (one of two energetically close stacking configurations; compare with Extended Data Fig. 8d). The projection of the latter two structures matches the experimental observations well. Panels **a**, **c**, **e** are top views, and panels **b**, **d**, **f** are side views along the dashed line given in the respective top view. C atoms and sp^2 bonds are grey, Li atoms are magenta. Solid black lines indicate the in-plane lattice constants of the Li crystals. The insets in **a**, **c**, **e** schematically show diffraction patterns associated with the respective structure (scale bars, 2 nm^{-1}), with first-order diffraction spots from the Li (C) lattice indicated in magenta (grey). The close-packed Li phase in **c**, **e** may assume any relative rotation angle with respect to the bilayer graphene lattice at virtually no additional energy cost (Extended Data Fig. 7b). A faint magenta circle indicates that the Li diffraction spots may therefore be rotated with respect to those of graphene. E_f is the energy required to take a Li atom from a bulk Li crystal and insert it between graphene sheets in the corresponding configuration. Contour plots in **b**, **d**, **f** represent the charge transfer between Li and graphene as compared to the isolated graphene and Li crystals. An increase in the electron density (negative charge) is shown blue, and a decrease in the electron density (positive charge) is shown red.

displays colour renditions of the charge probability distribution as compared to isolated graphene and Li crystals consisting of a single, double or triple layer. As can be seen from Fig. 3f in the triple layer case, charge transfer is noticeable only for the outer Li layers directly neighbouring a graphene sheet. Inner Li layers retain their metallic character as electronic charge is distributed between the Li^+ lattice sites. Extended Data Fig. 8c depicts the average charge transfer per Li atom, which drops as the close-packed Li phase gets thicker. Renormalizing by the number of atoms in the outermost Li layers only, we find a constant charge transfer of approximately 0.33 e^- per outer Li atom, irrespective of how many (metallic) Li layers (with nearly zero charge transfer) are packed in between. This is consistent with the observation that the resistivity ρ_{xx} measured during lithiation tends to saturate rather than to decrease progressively as more Li enters bilayer graphene (Fig. 1c). Note that although for C_6LiC_6 (Fig. 3a, b) a higher value of 0.85 e^- per Li atom for the charge transfer applies (resulting in an estimated density of transferred electrons of 2.7×10^{14} cm^{-2} per graphene sheet), the much denser arrangement of Li atoms in the close-packed structure yields a higher electron density per graphene sheet (4×10^{14} cm^{-2}) despite the smaller charge transfer value.

During lithiation, the close-packed Li phase grows laterally between the graphene sheets (Fig. 4a and Supplementary Video 1). Figure 4a displays a series of digital dark-field versions of the original images, where the three grains of varying in-plane orientation have each been coloured differently. Whereas the grain boundary between the central grain (green) and the lower right grain (blue) is rather sharp and stable, the boundary with the left grain (red) appears more fuzzy and mobile as their orientations nearly match. Regions of different thickness can be identified even within a single grain (Extended Data Fig. 4). The specimen is too thin to allow reliable extraction of its exact thickness t_{Li} from EELS (we typically obtain $t_{Li}/\lambda < 0.1$, where λ is the inelastic mean free path of about 125 nm for 80 keV electrons), but one may nonetheless determine relative variations (see Methods). When imaging an extremely thin slab of weakly scattering elements in a microscope with sufficient resolving power, the imaging contrast increases with increasing specimen thickness. This is quantified in Extended Data Fig. 4e, f, revealing that thinner parts of the close-packed Li phase tend to be located closer to its perimeter. At the leading edge, single atoms can be identified (Methods, Extended Data Fig. 1). From the image time series, a lateral growth rate of the order of 1 Å s^{-1} can be extracted. During delithiation, the close-packed Li phase disassembles and gradually disappears (Fig. 4b). Eventually, the pure bilayer graphene lattice remains behind. The degree of reversibility of this process is limited by the number of defects in the graphene lattice and their irreversible formation during prolonged imaging (electron irradiation). Given the combination of slow image acquisition (of the order of 1 s) and low sensitivity to light atoms, rapid diffusion of single Li ions that are not ordered remains concealed from the TEM observer. It is nonetheless present, both within and beyond the superdense phase, and is likely to be responsible for the initial abrupt change in resistivity ρ_{xx} during lithiation (Fig. 1c).

We note that the observed crystalline phase of Li proves stable only between intact graphene planes. When the incident electron energy exceeds the threshold for displacement damage of Li (about 20 keV), conditions are such that Li readily 'boils' under the electron beam[16,18]. The close-packed phase is volatile when imaged near bilayer edges or in the presence of a high density of defects in the graphene lattice. Constituents may escape from between graphene sheets via such edges or defects on electron-beam-induced melting of the crystalline phase. Likewise, we find that material having escaped from within and agglomerated on the outer bilayer surfaces near such defects quickly evaporates under electron beam irradiation. The protective encapsulation by two impermeable atomic sheets may thus be regarded a prerequisite for safe probing by TEM of the crystal formation therein, akin to the situation in graphene liquid cells[23].

Close-packing of Li intercalated between graphene sheets, as demonstrated here, results in a structure with a Li content greatly in excess of LiC_6. Although enhanced Li storage has been previously proposed to occur on the outside of graphene planes[24,25], the suggested atomistic configurations were contradictory and could neither be addressed nor verified by microscopic means. Other reports claiming the formation of nanocrystallites of close-packed Li during the lithiation of different carbon allotropes are scarce and not fully consistent[26,27]. At elevated temperatures, a configuration of Li similar to that which we report here may have been left unidentified in bulk graphite[28].

Because the energy cost for close-packing of Li in the van der Waals gap of bilayer graphene is very similar to that of forming C_6LiC_6

© 2018 Springer Nature Limited. All rights reserved.

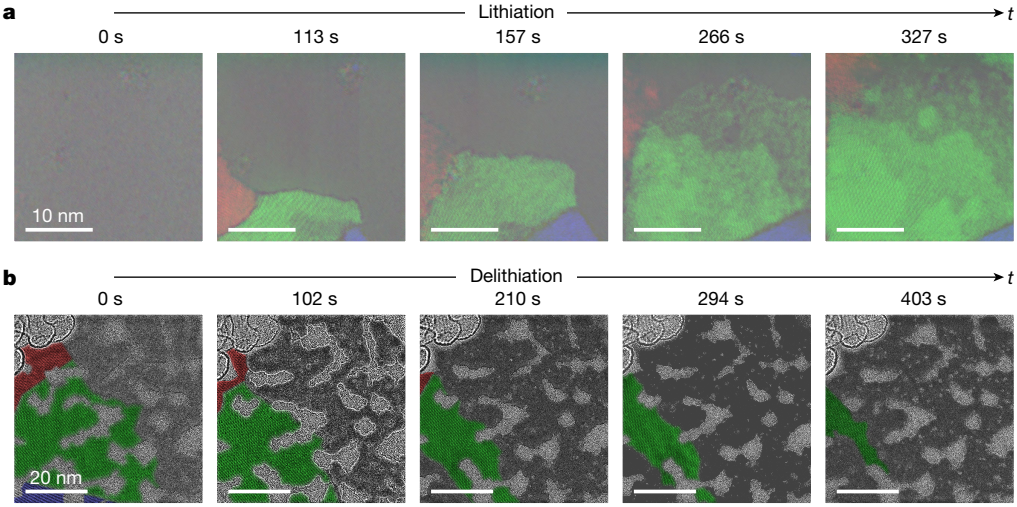


Fig. 4 | Li crystal growth between two graphene sheets. a, **b**, Time series of digital dark-field versions of the original images during lithiation (**a**) and (**b**) delithiation; times are shown at top of panels. Crystal grains of different in-plane orientation are coloured in red, green and blue. An increased amount of amorphous, immobilized residues (grey areas, essentially composed of carbon) are present within the field of view in **b**, an unavoidable consequence of extended electron beam illumination. In both **a** and **b** the interface with the electrolyte (not contained in the images) is oriented towards the bottom of the page.

(Fig. 3), close-packing may be the way the system accommodates a large amount of Li supplied within a short time. The activation energy for Li diffusion in bilayer graphene has been calculated (Methods, Extended Data Fig. 10). It points towards facile diffusion of Li between graphene sheets by the exchange mechanism, even in the presence of a close-packed Li phase. And although a graphene bilayer can be regarded as the structural unit of thicker graphite, its properties differ in many respects. Two atomic sheets may well spread more easily when isolated from their bulk crystal and thus render new types of intercalate ordering possible. Even if so, the stabilization of a superdense lithium phase comes as a surprise as its appearance has typically been reserved for extreme conditions[21] only. It is noteworthy that Li–graphite intercalation compounds only occupy a small region of the C–Li binary alloy phase diagram[29]. In the miscibility gap beyond LiC_6, alternative configurations may be available for storing larger amounts of lithium in layered carbons[30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0754-2.

Received: 8 January; Accepted: 15 October 2018;
Published online 26 November 2018.

1. Winter, M. & Besenhard, J. O. In *Handbook of Battery Materials* 2nd edn (eds Daniel, C. & Besenhard, J. O.) 433–478 (Wiley-VCH, Weinheim, 2011).
2. Kaskhedikar, N. A. & Maier, J. Lithium storage in carbon nanostructures. *Adv. Mater.* **21**, 2664–2680 (2009).
3. Zheng, H., Meng, Y. S. & Zhu, Y. Frontiers of in situ electron microscopy. *MRS Bull.* **40**, 12–18 (2015).
4. Qian, D., Ma, C., More, K. L., Meng, Y. S. & Chi, M. Advanced analytical electron microscopy for lithium-ion batteries. *NPG Asia Mater.* **7**, e193 (2015).
5. Liu, X. H. et al. In situ TEM experiments of electrochemical lithiation and delithiation of individual nanostructures. *Adv. Energy Mater.* **2**, 722–741 (2012).
6. Shao-Horn, Y., Croguennec, L., Delmas, C., Nelson, E. C. & O'Keefe, M. A. Atomic resolution of lithium ions in $LiCoO_2$. *Nat. Mater.* **2**, 464–467 (2003).
7. Oshima, Y. & Murakami, Y. (eds) *Microscopy* **66** (*Challenges for Lithium Detection special issue*), 1–61 (2017).
8. Reimer, L. & Kohl, H. *Transmission Electron Microscopy* (Springer Series in Optical Sciences, Springer, New York, 2008).
9. Linck, M. et al. Chromatic aberration correction for atomic resolution TEM imaging from 20 to 80 kV. *Phys. Rev. Lett.* **117**, 076101 (2016).
10. Enoki, T., Suzuki, M. & Endo, M. *Graphite Intercalation Compounds and Applications* (Oxford Univ. Press, New York, 2003).
11. Yu, Y. et al. Gate-tunable phase transitions in thin flakes of 1T-TaS_2. *Nat. Nanotechnol.* **10**, 270–276 (2015).
12. Bediako, D. K. et al. Heterointerface effects in the electrointercalation of van der Waals heterostructures. *Nature* **558**, 425–429 (2018).
13. Kühne, M. et al. Ultrafast lithium diffusion in bilayer graphene. *Nat. Nanotechnol.* **12**, 895–900 (2017).
14. Meyer, J. C. et al. Accurate measurement of electron beam induced displacement cross sections for single-layer graphene. *Phys. Rev. Lett.* **108**, 196102 (2012).
15. Mauchamp, V., Boucher, F., Ouvrand, G. & Moreau, P. Ab initio simulation of the electron energy-loss near-edge structures at the Li K edge in Li, Li_2O, and $LiMn_2O_4$. *Phys. Rev. B* **74**, 115106 (2006).
16. Wang, F. et al. Chemical distribution and bonding of lithium in intercalated graphite: identification with optimized electron energy loss spectroscopy. *ACS Nano* **5**, 1190–1197 (2011).
17. Lru, D.-R. & Williams, D. B. The electron-energy-loss spectrum of lithium metal. *Phil. Mag. B* **53**, L123–L128 (1986).
18. Hightower, A., Ahn, C. C., Fultz, B. & Rez, P. Electron energy-loss spectrometry on lithiated graphite. *Appl. Phys. Lett.* **77**, 238–240 (2000).
19. Liu, X. H. et al. In situ transmission electron microscopy of electrochemical lithiation, delithiation and deformation of individual graphene nanoribbons. *Carbon* **50**, 3836–3844 (2012).
20. Wyckoff, R. W. G. *Crystal Structures* Vol. 1, 2nd edn (Wiley & Sons, 1963).
21. Ackland, G. J. et al. Quantum and isotope effects in lithium metal. *Science* **356**, 1254–1259 (2017).
22. Sugawara, K., Kanetani, K., Sato, R. & Takahashi, T. Fabrication of Li-intercalated bilayer graphene. *AIP Adv.* **1**, 022103 (2011).
23. Yuk, J. M. et al. High-resolution EM of colloidal nanocrystal growth using graphene liquid cells. *Science* **336**, 61–64 (2012).
24. Sato, K., Noguchi, M., Demachi, A., Oki, N. & Endo, M. A mechanism of lithium storage in disordered carbons. *Science* **264**, 556–558 (1994).
25. Deschamps, M. & Yazami, R. Great reversible capacity of carbon lithium electrode in solid polymer electrolyte. *J. Power Sources* **68**, 236–238 (1997).
26. Wang, Q. et al. Investigation of lithium storage in bamboo-like CNTs by HRTEM. *J. Electrochem. Soc.* **150**, A1281–A1286 (2003).
27. Lee, B.-S. et al. Face-centered-cubic lithium crystals formed in mesopores of carbon nanofiber electrodes. *ACS Nano* **7**, 5801–5807 (2013).
28. Kambe, N. et al. Intercalate ordering in first stage graphite-lithium. *Mater. Sci. Eng.* **40**, 1–4 (1979).
29. Okamoto, H. In *Binary Alloy Phase Diagrams* Vol. 1, 2nd edn (eds Massalski, T. B., Okamoto, H., Subramanian, P. R. & Kacprzak, L.) **857** (ASM International, Ohio, 1990).
30. Lin, D. et al. Layered reduced graphene oxide with nanoscale interlayer gaps as a stable host for lithium metal anodes. *Nat. Nanotechnol.* **11**, 626–632 (2016).

Acknowledgements We acknowledge financial support from the Baden-Württemberg Stiftung gGmbH (project CT 5) and from the European Union Graphene Flagship. We are grateful to FEI/ThermoFisher Scientific for providing drawings and specifications of the NanoEx-i/v holder. F.B., J.B. and U.K. acknowledge funding from the German Research Foundation (DFG) and the Ministry of Science, Research and the Arts (MWK) of the federal state of Baden-Württemberg, Germany, in the frame of the SALVE project. A.V.K. thanks the Academy of Finland for support under project number 286279 and the DFG under project KR 4866/1-1. The theoretical study of Li diffusion (A.V.K.) was supported by the Russian Science Foundation (project identifier, 17-72-20223). We thank K. v. Klitzing for discussions and support and J. Popovic

© 2018 Springer Nature Limited. All rights reserved.

for useful comments on the manuscript. We acknowledge CSC Finland and PRACE (HLRS, Stuttgart, Germany) for generous grants of CPU time.

Reviewer information *Nature* thanks I. Honma and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.H.S. and U.K. composed the project. M.K. and S.F. fabricated samples and performed electrochemical measurements. F.B. performed the TEM and EELS experiments. M.G. and A.V.K. did the DFT calculations. J.B. helped with TEM and EELS experiments and did TEM imaging simulations. U.K. supervised the TEM work. D.S. contributed to the electrochemical design of the experiment. M.K. and J.H.S. wrote the manuscript and all authors contributed to it.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0754-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0754-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to U.K. or J.H.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Sample fabrication. Bulk natural graphite (NGS Naturgraphit) was exfoliated using adhesive tape[31] onto a sacrificial poly(methyl methacrylate) (PMMA) layer. A suitable bilayer graphene flake was selected with the help of an optical microscope. From the characteristic Raman scattering response, its bilayer nature[32] was verified. Using a dry-transfer method[33], the bilayer was placed over a hole in the Si_3N_4 membrane of a custom-made TEM sample carrier chip. A PMMA mask was patterned using electron-beam lithography followed by O_2-plasma etching in order to shape and isolate the desired flake. Multiple electrical contacts to the bilayer as well as the counter electrode were realized by standard lift-off techniques. In this work, samples with contacts made of either 60-nm-thick evaporated Ti or 60-nm-thick sputter-deposited Pt were used. The devices were annealed at 300 °C in a low-pressure (about 150 mbar) forming gas atmosphere. Inside an Ar-filled glovebox, 0.35 M lithium bis(trifluoromethane) sulfonimide (Li-TFSI) in polyethylene glycol methyl ether methacrylate:bisphenol A ethoxylate dimethacrylate (m-PEGMA:BEMA) w/w 3:7 electrolyte with added 2–4 wt% of 2-hydroxy-2-methylpropiophenone (HMPP) as photoinitiator was drop-casted on each sample, assuring only partial coverage of bilayer graphene (such that the sample area on the hole remained uncovered) and connecting it with the counter electrode. The electrolyte (see further details elsewhere[13,34]) was cured under ultraviolet radiation and capped by 200 nm evaporated SiO_x to prevent outgassing.

In situ TEM measurements. Measurements were performed inside the SALVE transmission electron microscope (see also http://www.salve-project.de) consisting of an FEI Titan Themis[3] column fitted with a CEOS aberration corrector. The corrector is a quadrupole-octupole corrector of modified Rose-Kuhn design that corrects for first order chromatic aberrations, fifth order axial geometric aberrations, and third order off-axial geometric aberrations[9]. The electron source is an FEI X-FEG Schottky type and the camera used is an FEI CETA 16M fibre coupled CMOS camera. Images were acquired using exposure times of 1 s at electron dose rates of about $(2–5) \times 10^6\ e^-\ nm^{-2}\ s^{-1}$ (4,096 by 4,096 pixels per frame). The camera pixel size at the magnification chosen for the lithiation time series was $(6.9\ pm)^2$ at an optical resolution of 70 pm. Local EELS spectra were taken with a Gatan Quantum ERS energy filter attached to the microscope. The microscope was operated at an electron acceleration voltage of 80 kV. From the full-width at half-maximum (FWHM) of the zero-loss peak we determine an energy resolution of about 0.7 eV. The sample space is kept at a base pressure of $p = 10^{-7}$ mbar, with the sample itself immersed in a static magnetic field of $B = 1.4$ T. We use an FEI NanoEx-i/v holder, providing eight electric feed-throughs to the sample using custom-made sample chip carriers. Four-terminal measurements of the electronic transport properties of uncovered bilayer graphene were done in situ using conventional lock-in techniques, with a low-frequency (13.33 Hz) a.c. excitation current of $I = 100$ nA applied across the bilayer. Lithiation (delithiation) was induced by applying a constant voltage $U_G = 5$ V ($U_G = 0$ V) to the metallic counter electrode using a source-measure unit. For atomic-resolution imaging, the graphene bilayer was cleaned in situ by current annealing[35] before the first lithiation. All experiments were performed at room temperature, with the sample exposed to the high vacuum environment of the transmission electron microscope.

TEM imaging resolution. The optical resolving power of the SALVE instrument[9], operated at 80 kV, is better than 0.08 nm. However, the true image resolution is element-specific, governed by the (scattering-angle-dependent) differential scattering cross-section. Li being a light element, its differential cross-section for scattering into high angles is small. Yet high-angle scattering is important for achieving high image resolution. In the Fourier transforms shown in Fig. 2a, d, f, one sees signals from the C lattice up to $(71\ pm)^{-1}$, indicating a nonlinear information transfer up to at least 0.07 nm. The highest Li lattice frequency in Fig. 2d, f is $(90\ pm)^{-1}$, indicating a nonlinear information transfer up to at least 0.09 nm. If we were imaging a strong scatterer or a thick specimen in a conventional C_s-corrected microscope, a conservative estimate for the upper limit of the linear information transfer (that determines the real imaging resolution) would be given by half the value of the highest observed frequency, that is, 0.14 nm for C and 0.18 nm for Li. However, first of all, we are dealing with an extremely weak scatterer and the sample is very thin, even in the context of TEM . Thus, the nonlinear information transfer portion is small. Second, the instrument used is chromatic-aberration-corrected in addition to the geometric aberration correction. Unlike in a C_s-only corrected microscope, the nonlinear information transfer is dampened in approximately the same way as the linear information transfer[36]. For these two reasons, our real imaging resolution in fact closely corresponds to the value of the highest frequency in the Fourier transforms, that is, about 0.07 nm for C and 0.09 nm for Li. These values are not only smaller than the observed in-plane lattice constant $a_{Li} \approx 3.1$ Å of the close-packed Li phase, but also smaller than the nearest possible distance between the projection (along the *c* axis) of Li atoms located in different layers of this phase, which is about 1.8 Å. They are even smaller than the distance between nearest-neighbouring C atoms in graphene (1.42 Å). True atomic image resolution of the C and Li lattices is therefore achieved.

TEM sensitivity. In our TEM images we may identify individual Li atoms where the crystalline Li phase appears thinnest, that is, at the leading edge. In Extended Data Fig. 1e we highlight single Li atoms, appearing as dark spots at the very front of the growing Li phase. These spots are not due to image delocalization, since they are non-periodic. In the given regime of weakly scattering elements and very few atoms, the image contrast simply adds up. Consequently, the contrast value of a single C atom (atomic number 6) is identical to that of two superposed Li atoms (atomic number 3). Vice versa, the contrast value of a single Li atom corresponds to half the value of a single C atom. By Fourier filtering the TEM images, the information of a perfect bilayer graphene lattice is subtracted from the images. At sites where one C atom is missing in one of the two graphene sheets, this procedure adds the contrast information of a single C atom, which may serve as a reference. Extended Data Fig. 1f, g displays two line profiles of the imaging contrast (Extended Data Fig. 1f left and right): one of these runs across two Li atoms at the leading edge, and the other is centred on such a reference site (Extended Data Fig. 1g left and right). The contrast value of each Li atom is indeed about half that corresponding to one C atom.

Image simulations were carried out using the open-source program QSTEM (http://qstem.org). To create atomistic models, standard tools like Materials Studio (Accelrys Inc.) as well as a custom-made program were used. The following simulation parameters determined from the experiments were applied: electron energy $E = 80$ kV, spherical aberration $C_S = -10$ μm, remaining residual focus spread $\Delta_{FS} = 0.5$ nm, image spread $\Delta_{IS} = 25$ pm, convergence angle $\alpha = 0.5$ mrad, and applied electron dose $2 \times 10^6\ e^-\ nm^{-2}$ using Poisson statistics. The focus was used as a free parameter. The SALVE TEM image simulations, summarized in Extended Data Fig. 2a–e, show that the detection limit of the instrument with the imaging parameters used is sufficient to resolve single Li atoms. The bilayer graphene lattice was removed from simulated images by Fourier filtering. Extended Data Fig. 2a, b displays the atomistic model: a hexagonal close-packed Li wedge (red) between two graphene sheets (cyan). To the left of the edge of the Li wedge, five additional Li atom rows were included. Furthermore, several vacancies were introduced to the Li lattice. In panels c–e the corresponding dose-limited TEM image simulations are shown. The Li wedge with graphene above and below does not directly reveal single Li atoms (Extended Data Fig. 2c). However, these atoms become clearly apparent in the filtered image in Extended Data Fig. 2d, as do the vacancies in the Li lattice. To verify the accuracy of the filtering procedure, a dose-limited TEM image simulation of the Li-wedge only, that is, without the graphene lattice (see Extended Data Fig. 2e), was performed. Both the single Li atoms and the vacancies are clearly visible, showing contrast similar to that in the filtered image (Extended Data Fig. 2d).

Fourier filtering. The original images shown in Fig. 2a–c, h are unfiltered; in these, the graphene bilayer lattice is atomically resolved. Additionally, we show images where the graphene lattice and the moiré effects have been filtered away (Fig. 2g, i); we cut out only sharp signals in Fourier space, and the cutout was very local. The signals of the Li phase are not affected at all. Note that neither a bandpass filter nor a Wiener filter was used, as these might have a slight impact on the atom positions. Nor did we use a filter that isolates the Li signals and may additionally enhance the apparent resolution of the image. The exact mask applied in Fourier space and the remaining Fourier transform are illustrated in Extended Data Fig. 3a and b, respectively. The filtered Fourier transform (Extended Data Fig. 3b) of the original image was back-transformed to real space (Fig. 2g). In the filtered Fourier transform (Extended Data Fig. 3b), the streaking (highlighted in yellow) resulting from the edges of the real image is very prominent in every spot. This streaking was not filtered out in order to avoid masking too-large portions of Fourier space. The streaking is responsible for the artefacts at the edges of the filtered image. Finally, the inhomogeneity of the illumination in the original image is gone in the filtered image because the central part of the Fourier transform was also cut out.

One effect of filtering out the bilayer graphene is that effectively a perfect lattice is subtracted from the image. Local defects in the graphene lattice thus become visible in the filtered image. For example, a vacancy in one graphene layer appears as a carbon atom with inverted contrast in the filtered image, as shown in Extended Data Fig. 1f–h.

Thickness determination. The observed close-packed Li phase reveals regions of different thickness. This is illustrated in Extended Data Fig. 4a, where major edges enclosing regions of approximately constant thickness have been marked. The two Fourier transforms identify regions of apparently different thickness but belonging to a single grain. EELS measurements of the Li phase during lithiation of a graphene bilayer have been performed. The 80 keV electron beam was spread over a region of several hundred nm^2 to minimize the areal dose. EEL spectra have been recorded without apparent detrimental effects of beam exposure on the Li

© 2018 Springer Nature Limited. All rights reserved.

phase. From the EELS data, we may extract an estimate for the thickness t_{Li} of the Li phase using:

$$\frac{t_{Li}}{\lambda} = \ln\left(\frac{I_t}{I_0}\right)$$

with I_t the integrated intensity of the total EEL spectrum and I_0 the integrated intensity of the direct beam. The inelastic mean free path $\lambda \approx 125$ nm for 80 keV electrons in Li was determined according to ref. [37]. For all EEL spectra acquired on lithiated bilayer graphene, we find $t_{Li}/\lambda < 0.1$, with typical values of $t_{Li} \approx 3$–4 nm. While this finding indeed points to a thin Li phase, it certainly does not represent its true thickness. The described method relies on the assumption that the integrated energy loss stems from bulk excitations, while surface excitations are entirely neglected. In the case at hand however, surface excitations are by far the dominant source of electron energy loss. Also, graphene surface excitations are contained in the spectrum, rendering extraction of information about the Li phase non-trivial. Therefore, the stated value represents an upper bound to the true thickness of the Li phase.

Knowing that the specimens are thin, the electron scattering cross-sections small, and the resolution of the SALVE TEM sufficient, relative variations in thickness of the close-packed Li phase may be determined as follows. Under these conditions, the contrast in the TEM images is expected to increase with increasing specimen thickness. Extended Data Fig. 4e is a reprint of Extended Data Fig. 4a, where green dashed lines demarcate boundaries between regions of approximately constant thickness within a single grain of the observed Li phase. From three regions that display different contrast (i, ii and iii in Extended Data Fig. 4e), we extract the histograms plotted in Extended Data Fig. 4f. The FWHM attests to the respective strength in contrast of these regions. For the conditions of the experiment, the larger the FWHM of the histogram the thicker (locally) the Li phase is. In the given case, the Li phase in region i is thinnest, while in region iii it is thickest.

Extended Data Fig. 2f, g displays SALVE TEM image simulations for Li wedges assuming two different close-packed Li structures: cubic and hexagonal. Extended Data Fig. 2a displays the atomistic models used, both in side view and in top view. The incident 80 keV electron beam is parallel to the normal of the layer planes. The corresponding image simulations are shown in Extended Data Fig. 2g for two different focus values (+6 nm overfocus and −6 nm underfocus; both at Lentzen conditions). Indeed, the tendency of increasing contrast for increasing number of close-packed layers for both systems is confirmed.

Low-loss EELS. Low-loss EELS data were recorded on graphene bilayers both before and during lithiation. Typical spectra are shown in Extended Data Fig. 5a. The dose rates were approximately $5 \times 10^6\ e^-\ nm^{-2}\ s^{-1}$ and the illuminated area was slightly larger than the imaged area. For the following, the optical mode was switched to diffraction with the central beam entering the entrance aperture of the spectrometer. Typical integration times were 10 s. Upon lithiation, the close-packed Li phase formed between graphene sheets. Concomitant with the Li-K edge, a pronounced new peak appeared in the low-loss region near 9 eV (Extended Data Fig. 5a). A similar feature was reported from EELS measurements on lithium metal and attributed to a Li plasmon mode[17,18]. We note the occasional observation of a weak shoulder near 30 eV, which can also be seen in Extended Data Fig. 5a. This shoulder has been attributed to oxidized lithium[17,18] and hence we do not exclude the presence of trace amounts of it within the probed area of the sample.

First-principles calculations. Density functional theory (DFT) calculations were carried out in the plane-wave basis and within the projector-augmented wave description of the core regions, as implemented in the VASP code[38,39]. The exchange-correlation functional proposed by Perdew, Burke and Ernzerhof[40] (PBE) and the van der Waals functional suggested by Björkman[41] were adopted for simulations of van der Waals-bonded 2D materials. An energy cut-off of 600 eV was used for the primitive cell calculations. For supercells we used 400 eV and 300 eV for the largest supercell with 624 atoms, corresponding to the 12×12 supercell of C_6LiC_6. The Brillouin zones of the primitive cell and supercells were sampled using $12 \times 12 \times 12$ (bulk lithium) and $2 \times 2 \times 1$ (2D supercells) Monkhorst–Pack grid points[42]. The maximum force on each atom is set to be less than 0.01 eV for optimized structures. The atomistic structures and the difference charge densities were illustrated using VESTA[43].

Various initial configurations of Li atoms between graphene sheets were created, and then the geometry was fully optimized (Extended Data Figs. 6–8). We considered both finite clusters of Li atoms and periodic commensurate Li–graphene systems. In order to reduce the lattice mismatch for periodic structures, the optimum size of the supercell and rotational angles between surfaces was found using the Virtual NanoLab software[44]. Finite Li clusters were encapsulated inside a 12×12 graphene bilayer supercell composed of 576 atoms. We simulated both AA and AB stacked graphene. The formation energy E_f is obtained from

$$E_f = \frac{E_{graphene+Li} - E_{graphene}}{n_{Li}} - \mu_{Li}$$

where $E_{graphene+Li}$ is the energy of the supercell containing bilayer graphene and Li atoms, $E_{graphene}$ is the energy of pristine bilayer graphene, n_{Li} is the total number of Li atoms and μ_{Li} is their chemical potential.

Electron energy loss near edge structure (ELNES). The ELNES simulations were performed using the all-electron full potential linear augmented plane wave (FLAPW) method with the TELNES program[45] as implemented in the WIEN2k package[46,47]. The calculations were all done within the generalized gradient approximation (GGA) using PBE exchange and correlation potential. The muffin-tin radii R_{mt} for Li and C atoms are 2.09 and 1.34 atomic units (a.u.), respectively. $R_{mt}K_{max}$ was fixed at 7.0 to determine the basis size. The full Brillouin zone was sampled by $5 \times 5 \times 1$ k points. The ELNES spectra were calculated using the double differential scattering cross-section (DDSCS) for the excitation of an atom by fast electrons and is given by the expression:

$$\frac{\partial^2}{\partial E \partial \Omega} = \frac{4\gamma^2}{a_0^2}\frac{k}{k_0}\frac{1}{Q^4}S(Q,E)$$

where a_0 is the Bohr radius, γ is the relativistic factor, k_0 and k are the electron wave vectors before and after interaction, respectively, and $Q = k_0 - k$ is the momentum transfer. In this approximation:

$$S(Q,E) = \sum_{i,f} |\langle i|e^{iQr}|f\rangle|^2 \delta(E + E_i - E_f)$$

for excitations from the initial state $|i\rangle$ with eigenvalue E_i to the final state $|f\rangle$ with eigenvalue E_f. The EELS calculations were performed using actual experimental conditions, such as beam energy, collection semi-angle, beam orientation and a spectrometer broadening of 0.75 eV. For comparison reasons, the simulated spectra were shifted to the major feature of the experimental spectrum (Extended Data Fig. 5b).

Li diffusion in bilayer graphene. In a first step, the diffusion barrier of a single Li atom between graphene sheets was calculated. For simplicity, only the relevant AB stacking configuration was considered. We obtained a barrier of 0.47 eV, in agreement with reported values for Li migration in graphite (0.4–0.6 eV, obtained using similar methods), see for example refs [48–51] and references therein.

In a second step, the diffusion barrier for lateral diffusion of a Li atom within two close-packed layers of Li encapsulated between two graphene sheets was calculated. To this end, an extra interstitial Li atom was considered as illustrated in Extended Data Fig. 10. The presence of the neighbouring graphene sheet forces the atom to take a position almost within the Li plane, in the middle of a Li triangle. In reality, the configuration gets distorted due to the interaction with the graphene layer. Note that the extra atom 'pushes' away the neighbouring Li atoms to make some space. The extra atom diffuses through the exchange mechanism with the nearest Li atom (see Extended Data Fig. 10c, d). One may think about this diffusion process as the motion of a Li–Li dimer. For its activation, we find a barrier lower than 0.3 eV. In reality, the value should depend on the exact position of the extra Li atom in the moiré pattern and correspondingly on the orientation of the Li lattice with respect to the graphene lattice. However, as the diffusion proceeds within the plane, we do not expect a strong dependence. It thus seems that at room temperature the Li atoms may easily move within the close-packed Li phase.

The outside of bilayer graphene in principle offers an alternative pathway for lateral diffusion of Li ions. Followed by vertical diffusion via defect sites, Li ions may thus potentially reach any place between the two graphene sheets. And even though Li diffusion on the outside of graphene sheets is expected to be fast[52], we did not observe a measurable efficiency of this diffusion pathway in previous work[13]. Despite a low theoretical barrier for Li migration on graphene's outside[51], of the order of 0.2 eV, it is energetically much more favourable for Li to reside between graphene sheets (the difference being about 0.6 eV). Unlike the case where Li is evaporated first on the outside of graphite followed by rapid lateral diffusion and intercalation[52], in our case Li ions from the electrolyte intercalate in between the graphene sheets followed by lateral diffusion within the sheets. This has been confirmed in the study in ref. [13].

Data availability

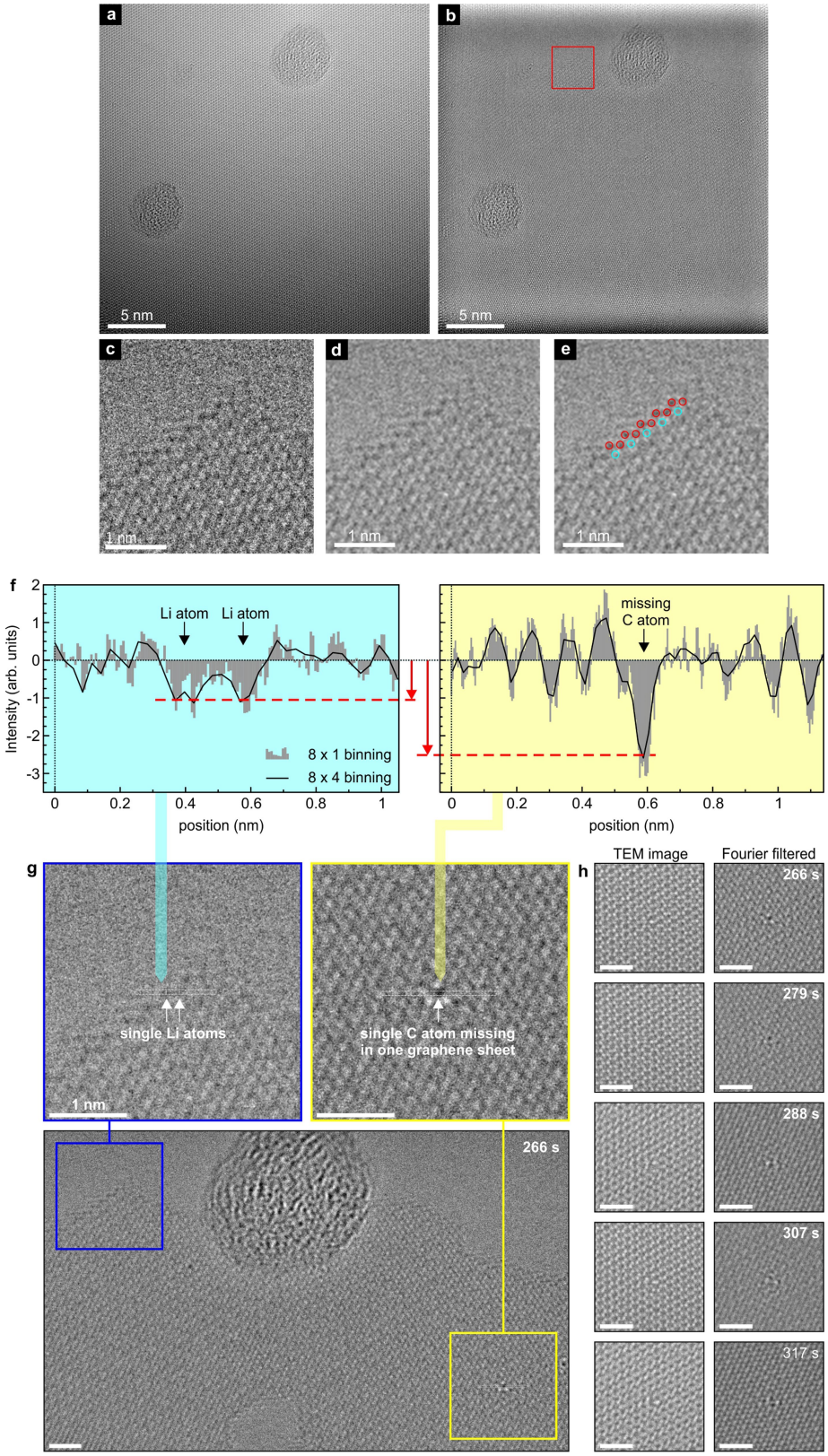
The data that support the findings of this study are available from the corresponding authors on request.

31. Novoselov, K. S. et al. Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004).

© 2018 Springer Nature Limited. All rights reserved.

32. Ferrari, A. C. et al. Raman spectrum of graphene and graphene layers. *Phys. Rev. Lett.* **97**, 187401 (2006).
33. Dean, C. R. et al. Boron nitride substrates for high-quality graphene electronics. *Nat. Nanotechnol.* **5**, 722–726 (2010).
34. Nair, J. R., Gerbaldi, C., Destro, M., Bongiovanni, R. & Penazzi, N. Methacrylic-based solid polymer electrolyte membranes for lithium-based batteries by a rapid UV-curing process. *React. Funct. Polym.* **71**, 409–416 (2011).
35. Moser, J., Barreiro, A. & Bachtold, A. Current-induced cleaning of graphene. *Appl. Phys. Lett.* **91**, 163513 (2007).
36. Haider, M., Hartel, P., Müller, H., Uhlemann, S. & Zach, J. Information transfer in a TEM corrected for spherical and chromatic aberration. *Microsc. Microanal.* **16**, 393–408 (2010).
37. Malis, T., Cheng, S. C. & Egerton, R. F. EELS log-ratio technique for specimen-thickness measurement in the TEM. *J. Electron Microsc. Tech.* **8**, 193–200 (1988).
38. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
39. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
40. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
41. Björkman, T. Van der Waals density functional for solids. *Phys. Rev. B* **86**, 165109 (2012).
42. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
43. Momma, K. & Izumi, F. VESTA: a three-dimensional visualization system for electronic and structural analysis. *J. Appl. Cryst.* **41**, 653–658 (2008).
44. Virtual NanoLab v. 2016.4 www.quantumwise.com (QuantumWise, 2016).
45. Nelhiebel, M. et al. Theory of orientation-sensitive near-edge fine-structure core-level spectroscopy. *Phys. Rev. B* **59**, 12807 (1999).
46. Blaha, P., Schwarz, K., Madsen, G. K. H., Kvaniscka, D. & Luitz, J. *WIEN2k, an Augmented Plane Wave + Local Orbitals Program for Calculating Crystal Properties* (ed. Schwarz, K.) (Techn. Universität Wien, Austria, 2001).
47. Schwarz, K. DFT calculations of solids with LAPW and WIEN2k. *J. Solid State Chem.* **176**, 319–328 (2003).
48. Persson, K., Hinuma, Y., Meng, Y. S., Van der Ven, A. & Ceder, G. Thermodynamic and kinetic properties of the Li-graphite system from first-principles calculations. *Phys. Rev. B* **82**, 125416 (2010).
49. Thinius, S., Islam, M. M., Heitjans, P. & Bredow, T. Theoretical study of Li migration in lithium–graphite intercalation compounds with dispersion-corrected DFT methods. *J. Phys. Chem. C* **118**, 2273–2280 (2014).
50. Ganesh, P. et al. Binding and diffusion of lithium in graphite: quantum Monte Carlo benchmarks and validation of van der Waals density functional methods. *J. Chem. Theory Comput.* **10**, 5318–5323 (2014).
51. Leggesse, E. G., Chen, C.-L. & Jiang, J.-C. Lithium diffusion in graphene and graphite: Effect of edge morphology. *Carbon* **103**, 209–216 (2016).
52. Mandeltort, L. & Yates, J. T. Rapid atomic Li surface diffusion and intercalation on graphite: a surface science study. *J. Phys. Chem. C* **116**, 24962–24967 (2012).

© 2018 Springer Nature Limited. All rights reserved.

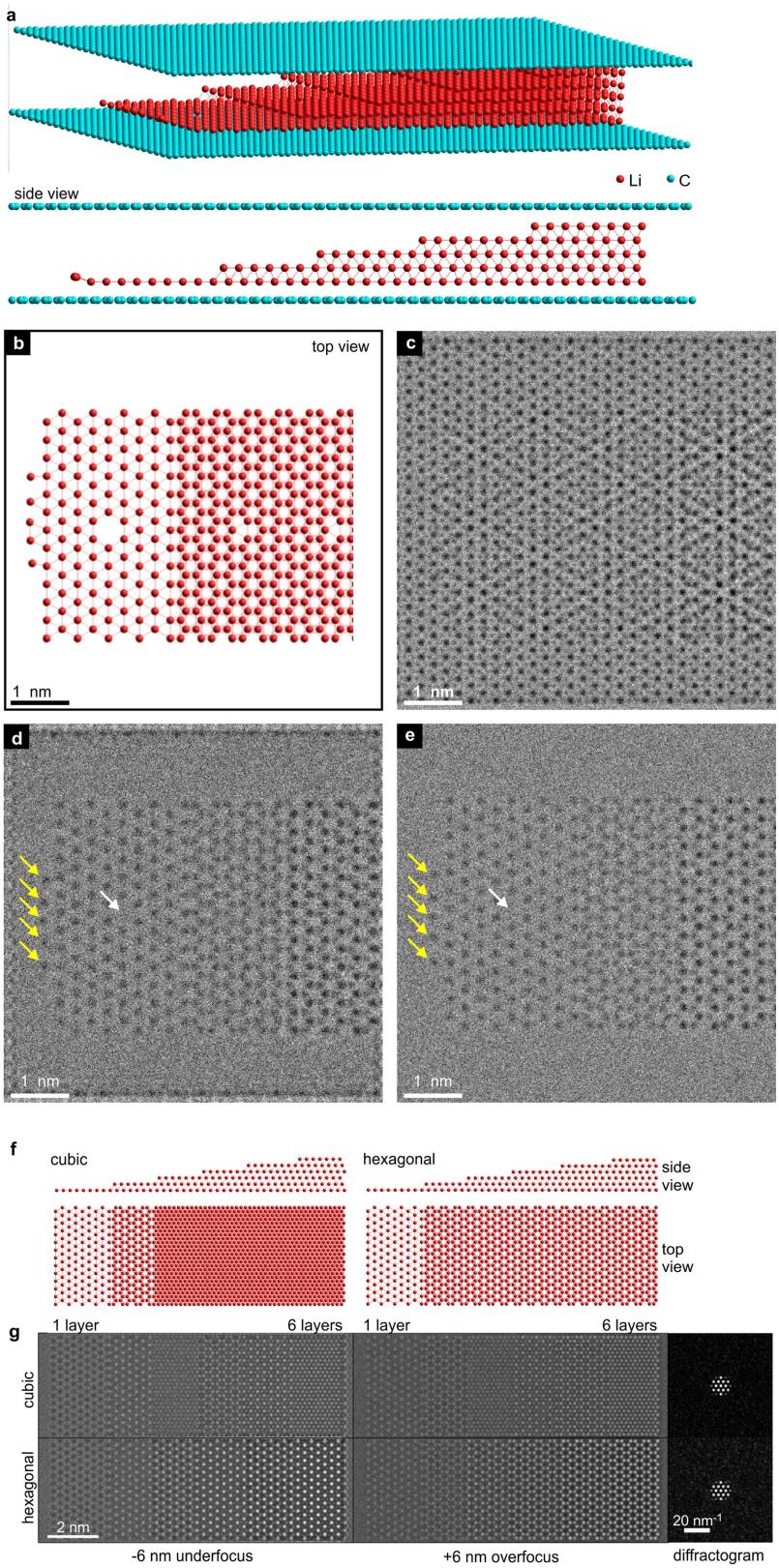


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 1 | Atomic resolution SALVE TEM images of lithium. a, Original TEM image (acquired between Fig. 2b, c). **b**, Same as **a**, but with graphene signals filtered out. **c**, Magnified view of area shown boxed in **b**. **d**, Slightly Gauss-filtered version of **c**. **e**, Same as **d**, but with red and blue circles indicating single Li atoms and symmetric positions without Li atoms, respectively. The latter show that the contrasts are not delocalization artefacts. **f**, Line profiles of the imaging contrast, centred on two neighbouring, individual Li atoms (left panel) and on the negative atomic contrast of a missing C atom in one graphene sheet, artificially inserted by the filtering procedure (right panel). The red arrows and red dashed lines indicate the signal intensity of the respective atomic species. **g**, Fourier-filtered version (bilayer graphene lattice removed) of a TEM image acquired during lithiation at $t = 266$ s (bottom image), between Fig. 2b, c. The top row shows two zoom-ins centred on locations where the line profiles in **f** have been extracted. **h**, Temporal evolution of the C vacancy in one of the two graphene sheets from which the line profile in **g** has been extracted at $t = 266$ s. Each row corresponds to data from a single time: left, magnified sections of our TEM images as acquired; right, the same sections after the removal of the graphene lattice by Fourier filtering. Scale bars, 1 nm.

© 2018 Springer Nature Limited. All rights reserved.

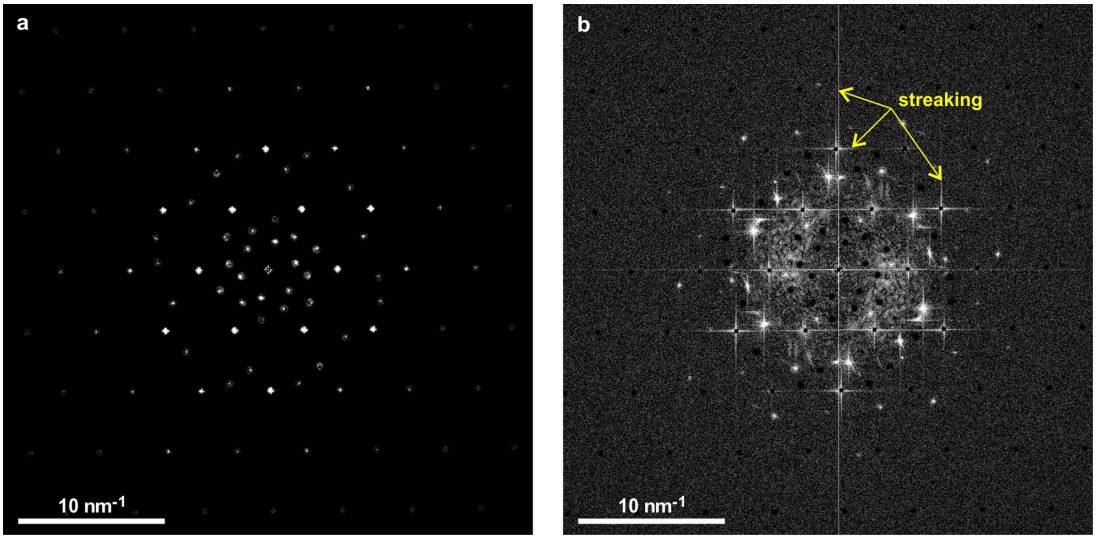


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

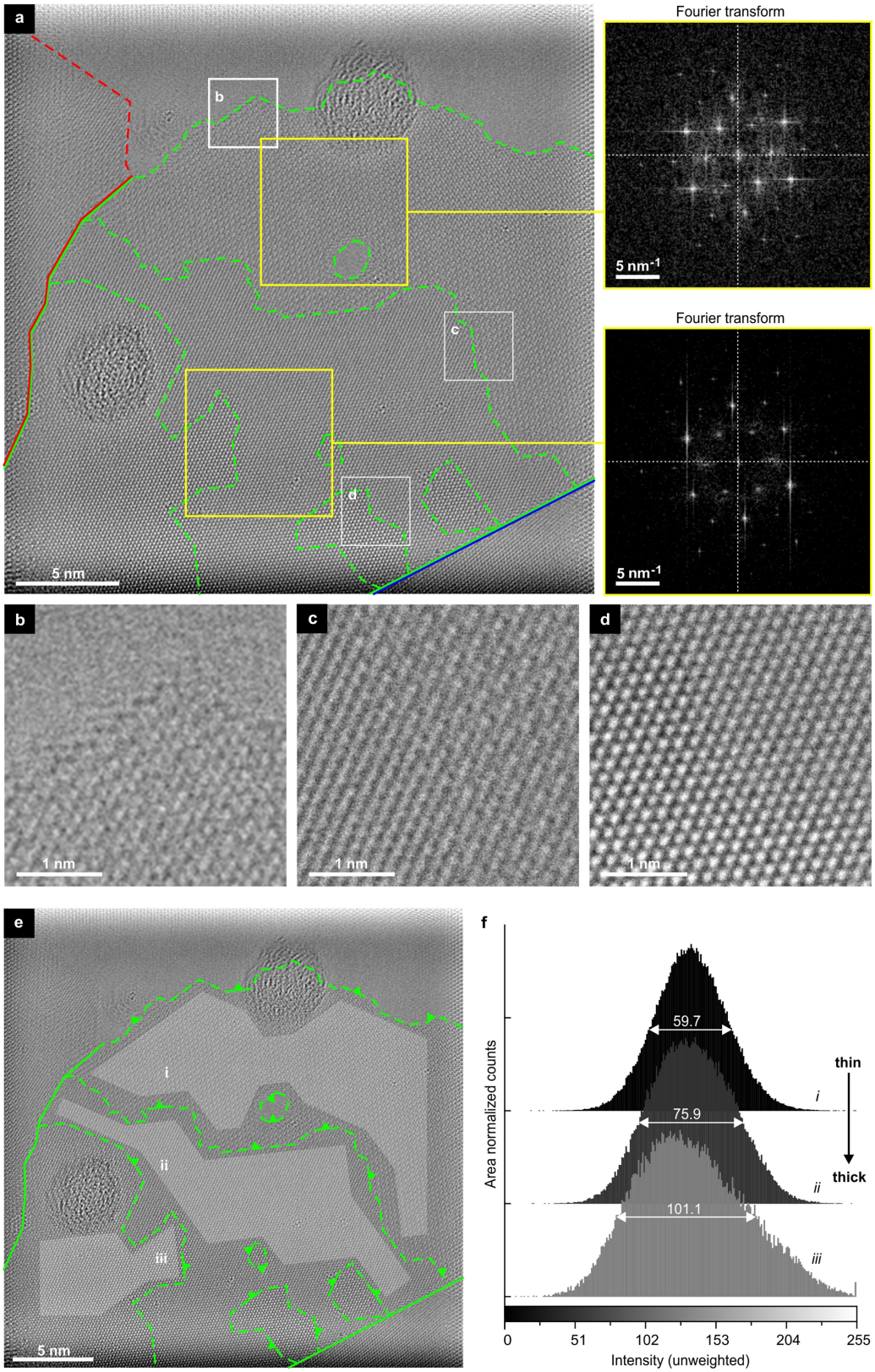
Extended Data Fig. 2 | 80 kV SALVE TEM image simulations. **a**, **b**, Atomistic model of a hexagonal close-packed Li wedge (red), with five additional single Li atom rows at its left edge as well as with single-atom vacancies, between two graphene sheets (cyan). Shown are a three-dimensional representation (**a**, top), a side view (**a**, bottom), and a top view of its thin front without C lattices (**b**). **c**–**e**, 80 kV dose-limited (applied dose 2×10^6 e^- nm^{-2}) TEM image simulations. **c**, Embedded in graphene. **d**, Fourier-filtered image with graphene removed. **e**, Unfiltered image simulated without graphene. In **d** and **e**, the yellow arrows mark single Li atoms and the white arrows point to a vacancy in the Li lattice. **f**, **g**, Image simulation of wedges of two different close-packed Li systems (cubic and hexagonal shown at left and right, respectively). **f**, Atomistic model used for the simulations, showing side and top views. The thickness gradually increases by one layer from left (one layer) to right (six layers). **g**, Image simulations for two different values of defocus, seen along [111] for cubic close-packed Li and along [0001] for hexagonal close-packed Li. The von Hann-filtered Fourier transforms on the right (diffractograms) are calculated for 4 layers at +6 nm overfocus. For the simulation the corresponding experimentally measured electron dose was applied.

© 2018 Springer Nature Limited. All rights reserved.



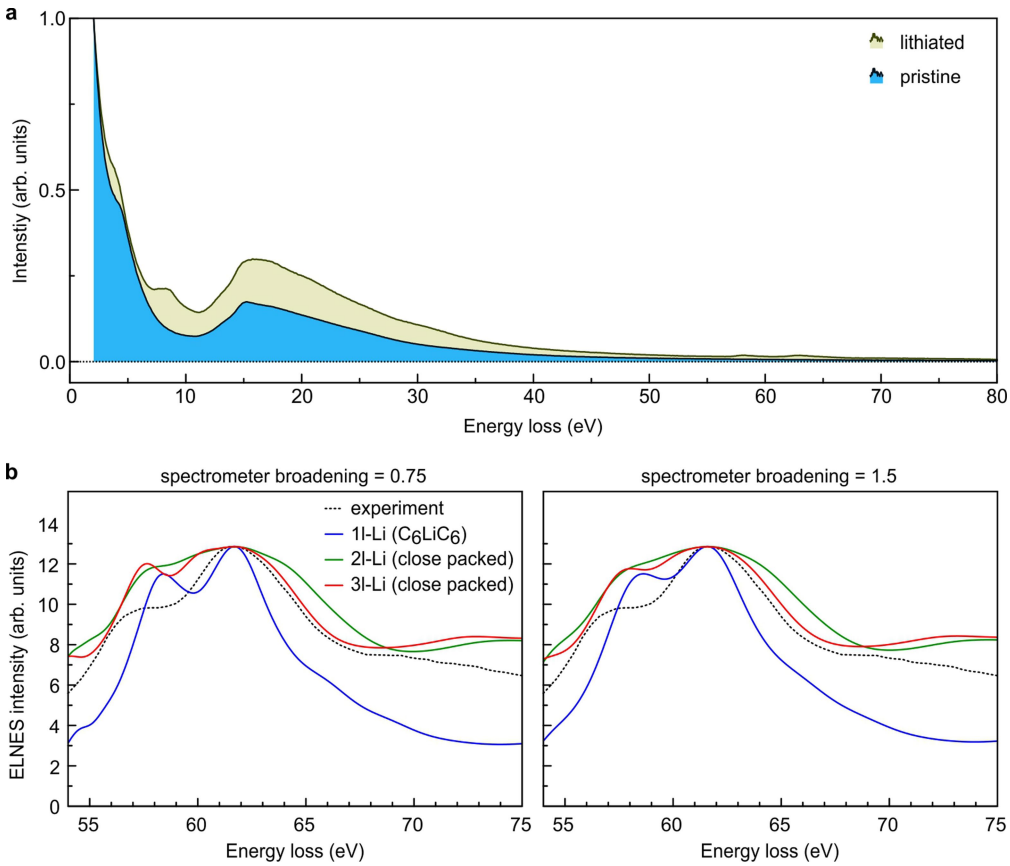
Extended Data Fig. 3 | Fourier filtering. Filtering mask for removing the graphene lattice and the moiré effects. **a**, Portion of the signal cut out by the applied mask. **b**, Remaining Fourier transform. The streaking (indicated by yellow arrows) results from the edges of the real image and was not Fourier-filtered to avoid masking too-large portions of Fourier space.

© 2018 Springer Nature Limited. All rights reserved.



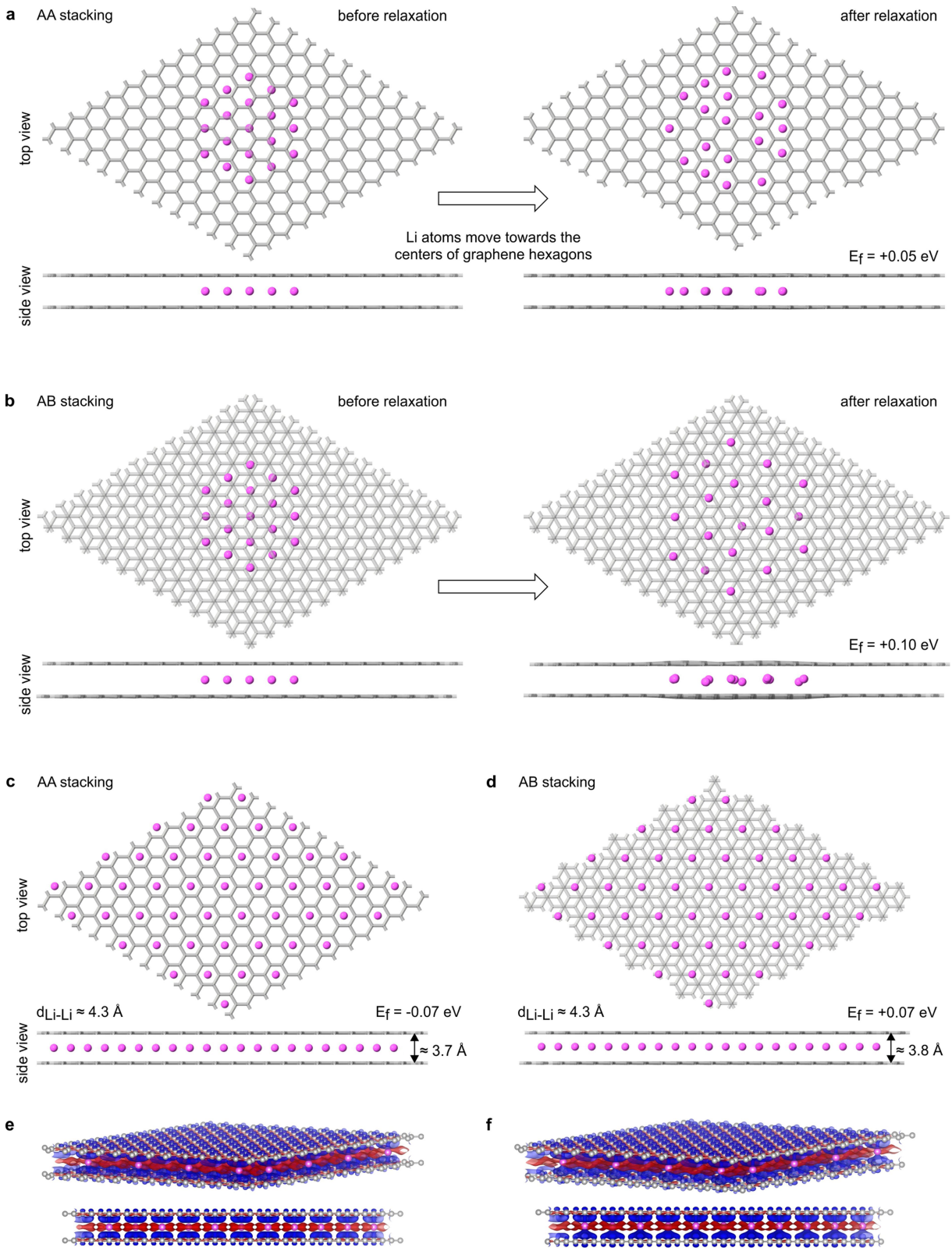
Extended Data Fig. 4 | Thickness determination of crystalline Li. **a**, Main panel, Fourier-filtered TEM image. Dashed lines demarcate major edges enclosing regions of the Li phase with approximately constant thickness. Two-coloured lines demarcate grain boundaries. Fourier transforms from two selected regions of different thickness (yellow boxes) are shown on the right. Signals from the Li phase in the Fourier transforms for both regions lying within one grain are identical. **b**–**d**, Magnified views of selected areas of the TEM image (white boxes) in **a**. **e**, **f**, Relative thickness determination from contrast quantification. **e**, Fourier-filtered TEM image (graphene lattice removed), identical to the main panel in **a**. Dashed lines demarcate major edges enclosing regions of the Li phase with approximately constant thickness. Arrowheads point outwards from thicker areas. **f**, Area-normalized intensity histograms acquired from the shaded regions in **e** that are labelled i, ii and iii. The full-widths at half-maximum (FWHMs) are indicated by white double-headed arrows.

© 2018 Springer Nature Limited. All rights reserved.



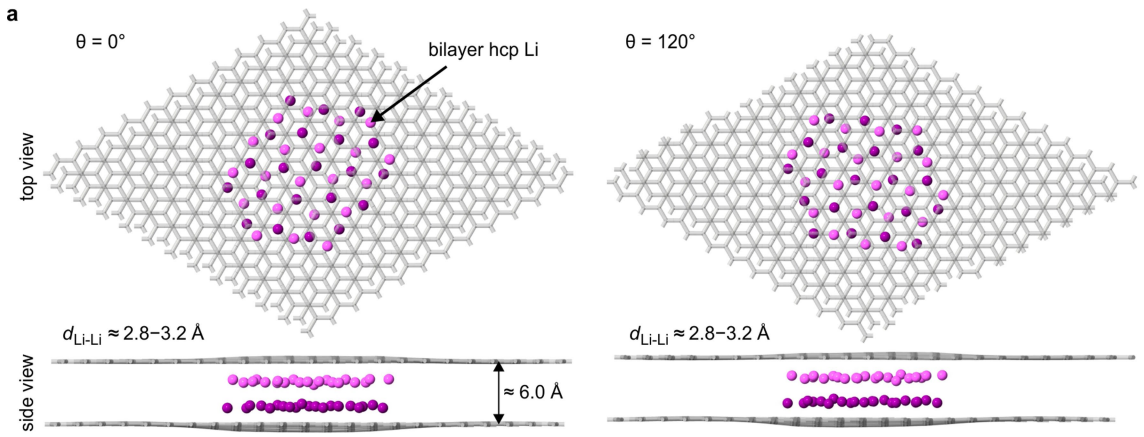
Extended Data Fig. 5 | Electron energy-loss spectroscopy. a, Low-loss EEL spectra of pristine (blue) and lithiated (yellow) bilayer graphene, with the close-packed Li phase present in the latter case. **b**, Calculated ELNES (electron energy loss near-edge structure; Methods) of the Li-K edge for bilayer graphene containing 1, 2 and 3 Li layers compared to the experimental ELNES. The spectrometer broadening is taken into account by convoluting the result with a Gaussian function. Two different broadenings have been considered (left and right panels). The spectrometer broadening is given as the FWHM of the respective Gaussian function.

© 2018 Springer Nature Limited. All rights reserved.

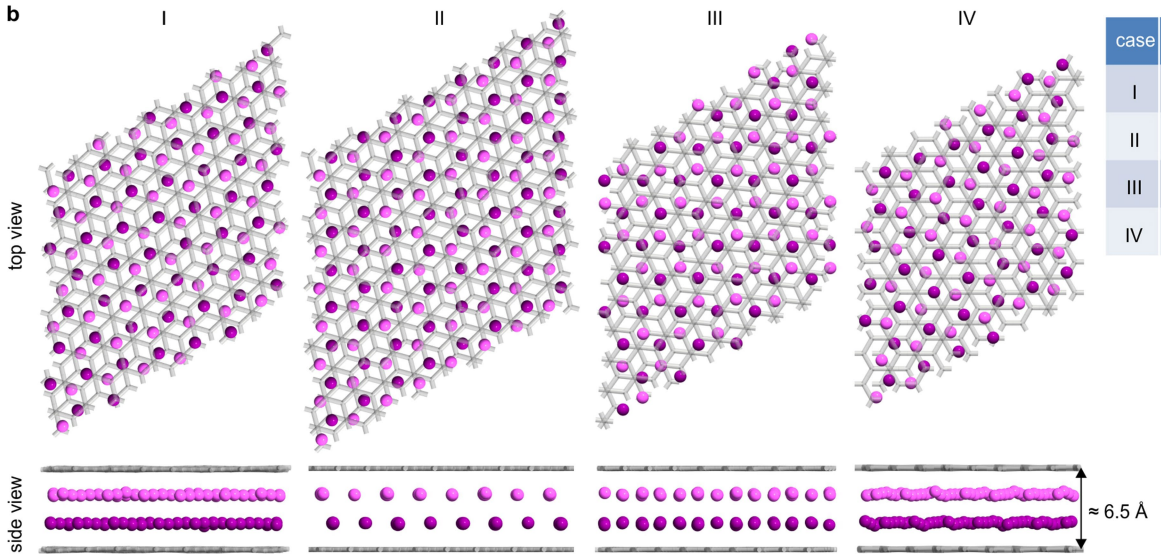


Extended Data Fig. 6 | Atomic configurations and energetics of a single layer of Li atoms between two graphene sheets. a, **b**, Structural evolution of a finite single-layer cluster of Li atoms between two graphene sheets with different stacking: AA-stacking (**a**) and AB-stacking (**b**). Left and right structures are before and after relaxation (that is, energy minimization); top and bottom are top and side views, respectively. The relaxation gives rise to the formation of a system with a geometry close to the C_6LiC_6 configuration. Note that the configurations correspond to a local, not global, energy minimum. **c**, **d**, The periodic C_6LiC_6 configuration with Li atoms arranged in a commensurate $(\sqrt{3} \times \sqrt{3})R30^\circ$ superstructure between graphene sheets for AA stacking (**c**) and AB stacking (**d**). $d_{\text{Li-Li}}$ refers to the separation of Li atoms. Double-headed arrows indicate the spacing between graphene sheets. **e**, **f**, Electron density difference between the combined system and its isolated parts. Red colour corresponds to a decrease in the electron density, blue to an increase. Charge transfer between Li and graphene (with an average value of 0.85 electrons per Li atom) is clearly observable.

© 2018 Springer Nature Limited. All rights reserved.



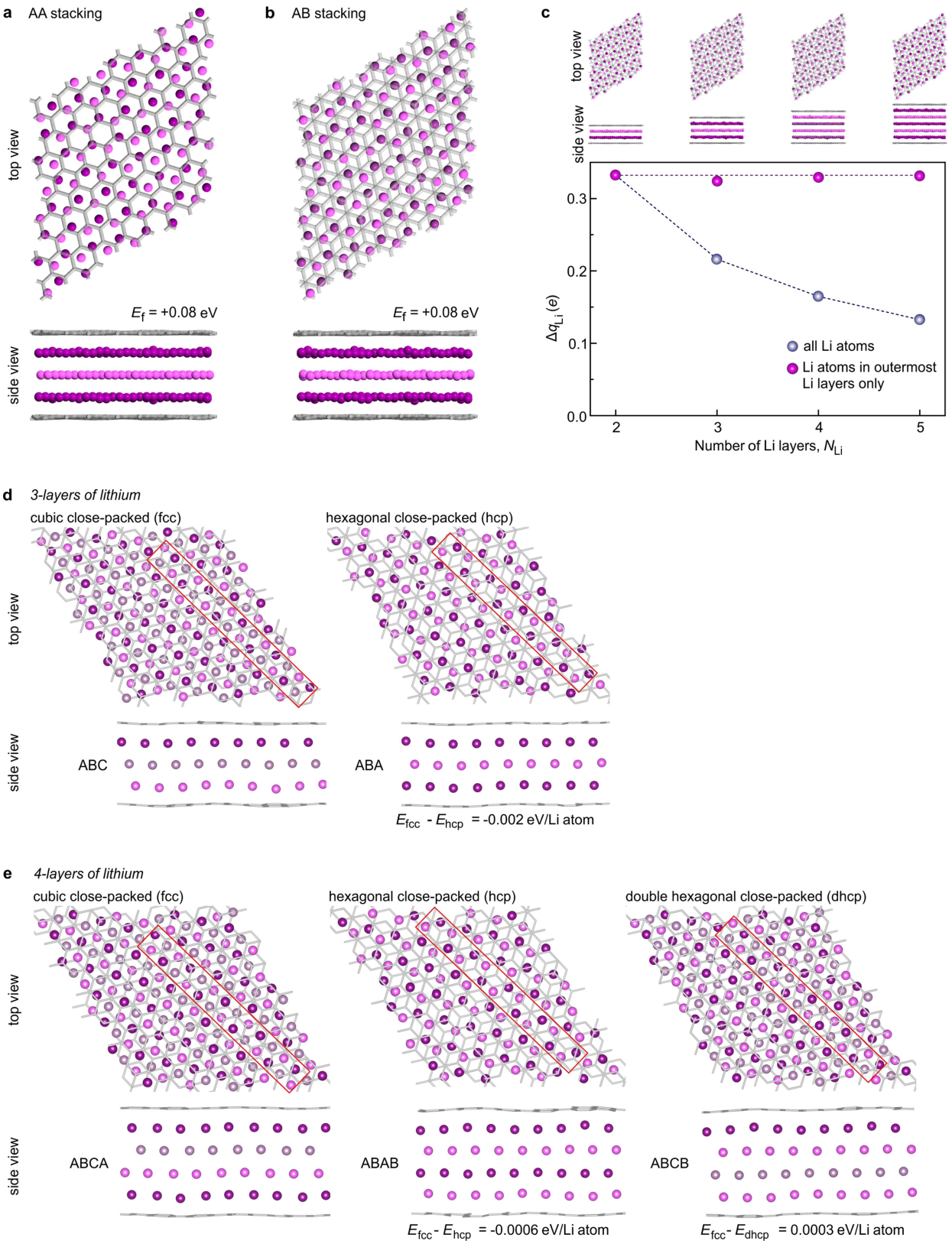
θ (°)	E_f (eV)
0	+0.31
60	+0.32
120	+0.31
180	+0.31
270	+0.34



case	θ (°)	strain	E_f (eV)
I	3.1	0.34%	+0.09
II	15.3	0.26%	+0.13
III	155.8	0.34%	+0.14
IV	172.4	0.34%	+0.13

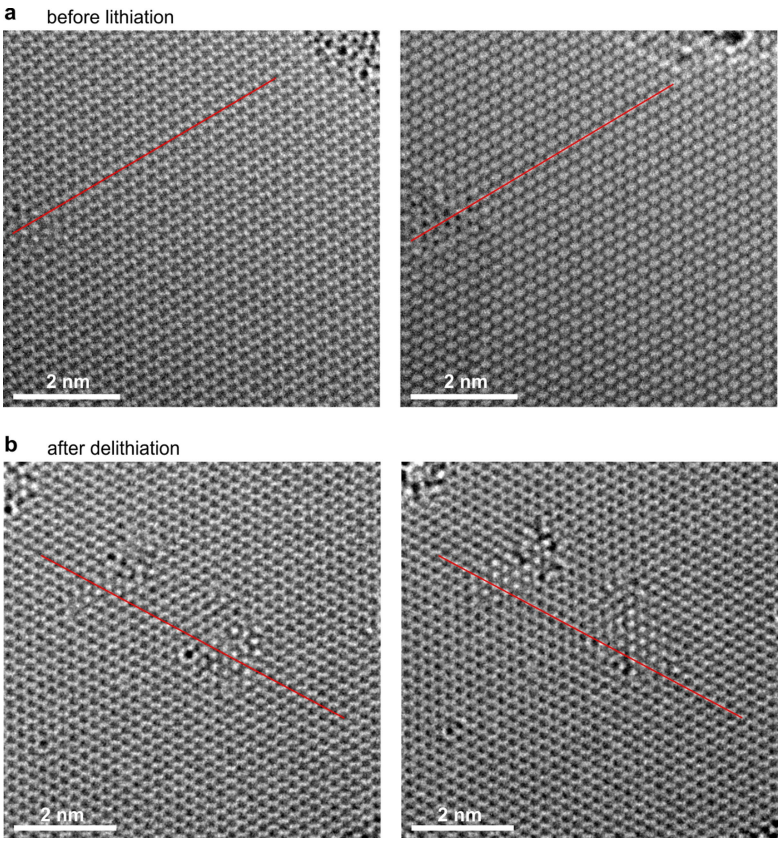
Extended Data Fig. 7 | Atomic configurations and energetics of Li bilayers between two graphene sheets. a, The geometric arrangement of the atoms of a finite bilayer cluster of Li atoms encapsulated between two graphene sheets after energy minimization for two different rotation angles θ between the Li and C lattices. The original configuration of the cluster was the perfect h.c.p. lattice. The structure is largely preserved during the relaxation. The distance between the Li atoms is denoted as d_{Li-Li}. A very weak dependence on the angle between graphene and h.c.p. lattice is found, as shown in the table at right. **b**, Atomic configurations and energetics of the infinite commensurate Li bilayer h.c.p./graphene structure and the dependence of the energy on the orientation angle θ between surfaces. Very weak dependence on the angle between graphene and the h.c.p. Li lattice is found. A small amount of strain introduced into the system to make the graphene and Li lattices commensurate affects the results by no more than 0.01 eV, as evident from the tables presenting E_f for different θ.

© 2018 Springer Nature Limited. All rights reserved.



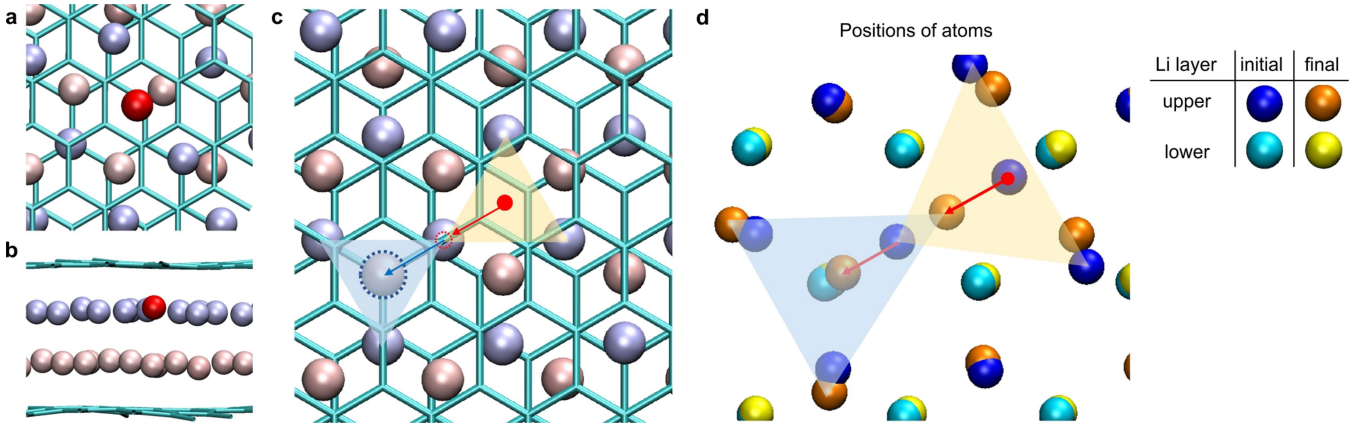
Extended Data Fig. 8 | Atomic configurations and energetics of Li multilayers between two graphene sheets. a, **b**, Atomic configurations and energetics of the infinite commensurate Li trilayer with h.c.p. structure between two graphene sheets for AA stacking (**a**) and AB stacking (**b**). **c**, Main panel, charge transfer from Li to graphene as a function of the number of close-packed Li layers N_{Li} between two graphene sheets. The corresponding atomic configurations are shown above the plot. The blue values are obtained as $\Delta q_{Li} = q_0 - q_{Li}$, where q_{Li} is the charge of Li after intercalation and q_0 is the charge of the isolated Li atom. Since the charge transfer is relevant for outermost Li layers only, we also plot the full charge transfer renormalized by the number of Li atoms in these outermost Li layers without considering the other Li layers (purple). **d**, **e**, Energy difference between different possible close-packed configurations (stacking orders), calculated for three layers of Li between two graphene sheets (**d**) and four layers of Li between two graphene sheets (**e**). To illustrate the stacking order, below each top view we include a side view of atoms within the red rectangle. The energy differences with respect to each f.c.c. case are stated below the side views.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Registry of graphene layers. a, **b**, Successive SALVE TEM images at different defocus values before lithiation (**a**) and after delithiation (**b**). The red lines are guides for the eye. We do not observe a change in registry of the two graphene sheets when comparing TEM images of the graphene lattice before lithiation and after delithiation. The registry can be checked at sites where one of the two graphene sheets features a moderately big hole. The stacking is AB without any sign of change throughout the experiment.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Lateral diffusion of a Li atom inside the close-packed Li system confined between two graphene sheets. a, **b**, Schematic atomistic configuration (top and side view). The red sphere represents the extra interstitial Li atom. **c**, Schematic of the diffusion process. **d**, The actual initial/final positions of atoms. The yellow and blue transparent triangles in **c** and **d** mark the initial and final configurations of the interstitial atom with regard to the nearest Li atoms in the undistorted (**c**) and the optimized (**d**) structures. The red and blue arrows connect the initial and final positions of the diffusing atoms.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0755-1

Catalytic deracemization of chiral allenes by sensitized excitation with visible light

Alena Hölzl-Hobmeier[1], Andreas Bauer[1], Alexandre Vieira Silva[1], Stefan M. Huber[2], Christoph Bannwarth[3] & Thorsten Bach[1]*

Chiral compounds exist as enantiomers that are non-superimposable mirror images of each other. Owing to the importance of enantiomerically pure chiral compounds[1]—for example, as active pharmaceutical ingredients—separation of racemates (1:1 mixtures of enantiomers) is extensively performed[2]. Frequently, however, only a single enantiomeric form of a chiral compound is required, which raises the question of how a racemate can be selectively converted into a single enantiomer. Such a deracemization[3] process is entropically disfavoured and cannot be performed by a conventional catalyst in solution. Here we show that it is possible to photochemically deracemize chiral compounds with high enantioselectivity using irradiation with visible light (wavelength of 420 nanometres) in the presence of catalytic quantities (2.5 mole per cent) of a chiral sensitizer. We converted an array of 17 chiral racemic allenes into the respective single enantiomers with 89 to 97 per cent enantiomeric excess. The sensitizer is postulated to operate by triplet energy transfer to the allene, with different energy-transfer efficiencies for the two enantiomers. It thus serves as a unidirectional catalyst that converts one enantiomer but not the other, and the decrease in entropy is compensated by light energy. Photochemical deracemization enables the direct formation of enantiopure materials from a racemic mixture of the same compound, providing a novel approach to the challenge of creating asymmetry.

Owing to the energetic degeneracy of enantiomers, an equilibrium in which either one of the enantiomers **A** or *ent*-**A** prevails (Fig. 1a) cannot be established either by a chiral or by an achiral catalyst (cat. I). However, there is ample precedence for a two-step process in which the two enantiomers are kept in a 50:50 equilibrium and one of them reacts in a second catalytic reaction in the presence of a chiral catalyst (cat. II), frequently an enzyme (dynamic kinetic resolution)[4]. Although thermally impossible, a photochemical equilibrium in favour of one enantiomer, **A** or *ent*-**A**, is potentially feasible in the presence of a chiral catalyst (cat. III) by sensitization (Fig. 1b). Previous work in this field has been mainly devoted to the isomerisation of (*E*)-double bonds in planar chiral alkenes by singlet energy transfer[5,6]. However, this process involves the formation of isomeric achiral (*Z*)-alkenes and thus occurs in low yields. In an alternative scenario, compounds that exist exclusively as enantiomers **A** and *ent*-**A** in the ground state, could be deracemized via triplet energy transfer from a chiral sensitizer. Allenes (1,2-propadienes) with an axis of chirality are known to undergo a configuration switch upon triplet-sensitized excitation[7] and are suitable substrates for photochemical deracemization. Their isomerisation occurs via an achiral planar triplet intermediate[8], which decays to enantiomers **A** and *ent*-**A**. For the deracemization of penta-2,3-diene, a low enantioselectivity of 3.4% enantiomeric excess (e.e.) was reported by employing a superstoichiometric loading (120 mol%) of a chiral sensitizer[9].

In preliminary experiments, we studied the chiral allene lactams **1a** and *ent*-**1a** (Fig. 1c), which we isolated as by-products in a photochemical reaction. An efficient three-step route towards a racemic mixture *rac*-**1a** was developed from *N*-(4′-methoxybenzyl)-3-iodo-2-piperidinone (see Supplementary Information, pages 6–11). The sensitization experiments were performed with the chiral thioxanthone **2**[10], whose triplet energy was determined as 263 kJ mol^{-1} (±3 kJ mol^{-1}; for details regarding uncertainties see Supplementary Information pages 76, 77) at 77 K (in trifluorotoluene).

The triplet energy of allene **1a** could not be determined because the compound did not display any detectable phosphorescence. Previous work on the parent allene[11] suggested, however, that the triplet state of allenes may be accessible by sensitization. Initial irradiation experiments were performed at a wavelength of $\lambda = 420$ nm with a racemic mixture of *rac*-**1a** (**1a**/*ent*-**1a** = 50/50, 0% e.e.) and 5 mol% of **2** in trifluorotoluene as the solvent at ambient temperature (Fig. 2). Remarkably, there was a rapid and almost complete deracemization (entries 1–4 in Fig. 2) and a photostationary state was reached after one hour (entry 3). Further irradiation did not improve the enantioselectivity (entry 4). After some optimization (see Supplementary Information, page 30), it was found that a substrate concentration of 10 mM and a catalyst loading of 2.5 mol% was ideal to achieve an almost complete deracemization at low catalyst loading (entry 5). In the less-polar solvent benzene the enantioselectivity was even higher (entry 6), but benzene does not allow a reaction at lower temperature owing to its high melting point. In this regard, the fact that deracemization in acetonitrile

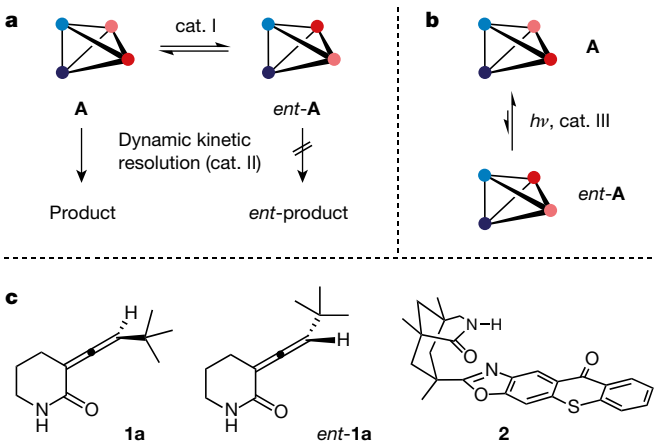
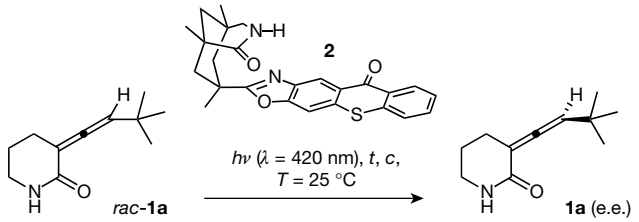


Fig. 1 | Reactions of generic chiral compound A and *ent*-A and structures of compounds 1, *ent*-1a and 2. a, A catalyst (cat. I) can under thermal conditions establish only a 50:50 equilibrium (racemate) of enantiomeric compounds **A** and *ent*-**A**. In a dynamic kinetic resolution, a second catalyst (cat. II) can lead to the desired formation of a single enantiomeric product. **b**, Upon irradiation (light energy, $h\nu$; h, Planck constant; ν, frequency) in the presence of a chiral sensitizer (cat. III) a photostationary state can be reached in which enantiomer **A** is formed at the expense of enantiomer *ent*-**A**. **c**, Allenes such as compounds **1a** and *ent*-**1a** are chiral compounds that can be photochemically interconverted. Compound **2** can act as a chiral catalyst by triplet sensitization.

[1]Department Chemie and Catalysis Research Center (CRC), Technische Universität München, Garching, Germany. [2]Fakultät für Chemie und Biochemie, Ruhr-Universität Bochum, Bochum, Germany. [3]Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Universität Bonn, Bonn, Germany. *e-mail: thorsten.bach@ch.tum.de

© 2018 Springer Nature Limited. All rights reserved.



Entry	c (mM)	Solvent	2 (mol%)	t (h)	e.e. (%)
1	5.0	$PhCF_3$	5.0	0.25	72
2	5.0	$PhCF_3$	5.0	0.5	88
3	5.0	$PhCF_3$	5.0	1	94
4	5.0	$PhCF_3$	5.0	4	94
5	10.0	$PhCF_3$	2.5	4	95
6	10.0	PhH	2.5	4	97
7	10.0	MeCN	2.5	4	95
8	10.0	MeOH	2.5	4	10

Fig. 2 | Optimization of reaction conditions. Top, conversion of a racemic mixture, *rac*-**1a**, of compounds **1a** and *ent*-**1a** into a major enantiomer, **1a**, upon irradiation in the presence of chiral sensitizer **2**. Bottom, reaction conditions for *rac*-**1a** (0% e.e.) in the given solvent under irradiation ($\lambda = 420$ nm) at room temperature ($T = 25$ °C). *c* denotes the substrate concentration and *t* is the reaction time. The enantiomeric excess (e.e.), (**1a** − *ent*-**1a**)/(**1a** + *ent*-**1a**), is determined by high-performance liquid chromatography analysis on a chiral stationary phase.

was equally successful was a surprising but pleasing discovery (entry 7). As expected, deracemization was not successful in a solvent in which hydrogen bonding between **1a** and **2** is completely precluded (entry 8).

The absolute configuration of the major enantiomer was determined by its circular dichroism spectrum, which matched the calculated spectrum of **1a**. The compound was levorotatory and showed a negative Cotton effect at $\lambda = 230$ nm (see Supplementary Information, pages 31–37). Other racemic allenes *rac*-**1** were subjected to the optimized reaction conditions (entry 7) at ambient temperature (Fig. 3a). Under these conditions, allenes with a primary, secondary or tertiary alkyl group at the endocyclic allene carbon atom gave consistently very high enantioselectivities (91%–97% e.e.; products **1a**–**1d**, **1f**–**1k**). Yields were also high (71%–98%) and in some cases the allene could be even quantitatively isolated. The 3-phenylpropyl-substituted allene **1e** turned out to be unstable under the applied irradiation conditions, resulting in a lower yield and slightly diminished enantioselectivity. It turned out that a decrease in reaction temperature to −40 °C led to an increase in enantioselectivity, and these conditions (Fig. 3b) were used for the sensitive methyl-substituted allene **1l** and for some functionalized allenes, **1m**–**1q**. For example, the 3-chloropropyl-substituted allene **1n** could be isolated at 87% yield and with 92% e.e. whereas the enantioselectivity at ambient temperature was only 86% e.e.

In line with previous work[12], it was shown that the allenes are configurationally stable if irradiated at $\lambda = 420$ nm in the absence of a sensitizer. Consequently, the observed deracemization seems to be linked to energy transfer from photoexcited thioxanthone to the two enantiomeric forms of the allene. This energy transfer can occur intermolecularly or intramolecularly in a complex of thioxanthone **2** with the allene, and the interconversion between **1a** and *ent*-**1a** proceeds via a planar triplet intermediate. In Fig. 4a, the two complexes are shown for allenes **1a** and *ent*-**1a**. Optimized structures for the complexes were obtained from density functional theory (DFT) calculations (see Supplementary Information, pages 38–45). Most notable is the fact that the distance of the allene double bond to the chromophore is considerably different for the two diastereomeric complexes. The distance between the carbonyl carbon atom of the thioxanthone and the terminal endocyclic carbon atom of the allene is 363 pm in complex **2**·*ent*-**1a** and 510 pm

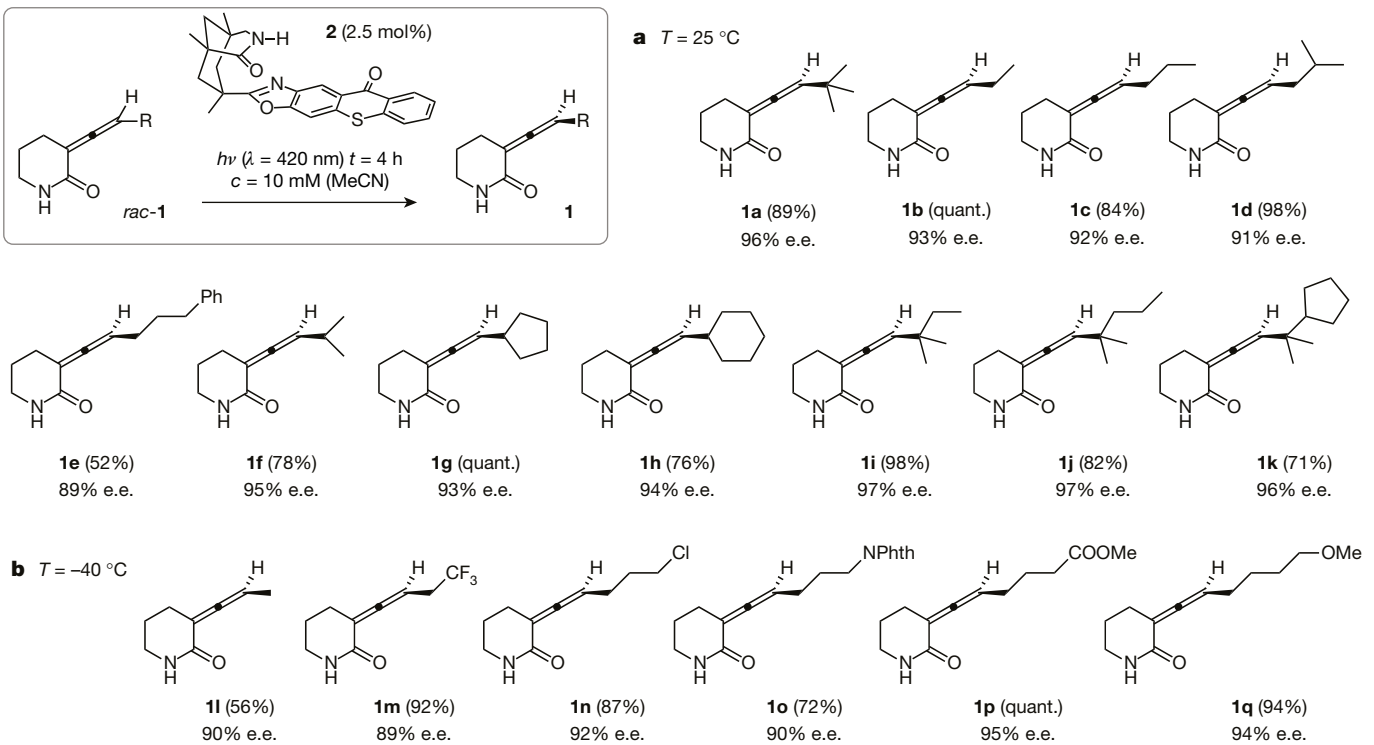


Fig. 3 | Deracemization of chiral allenes *rac*-1 by sensitized irradiation with visible light ($\lambda = 420$ nm). Inset, deracemization reaction scheme (*c*, concentration; *t*, reaction time; *T*, temperature). **a**, Allenes that are substituted with primary, secondary, or tertiary alkyl groups at the exocyclic allene carbon atom were irradiated at ambient temperature (25 °C). Yields refer to isolated product, and 'quant.' denotes quantitatively isolated allene. **b**, Reactions at −40 °C were performed in a cryostated apparatus with allenes that are methyl-substituted or that carry functional primary alkyl groups (Phth, phthaloyl) at the exocyclic allene carbon atom. Yields refer to isolated product.

© 2018 Springer Nature Limited. All rights reserved.

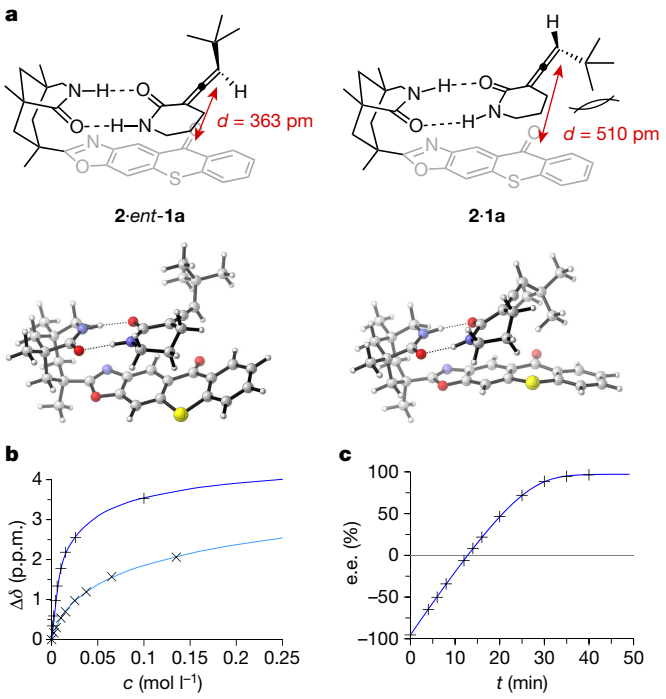


Fig. 4 | Mechanistic description of deracemization on the basis of DFT calculations, association constants and quantum yields. a, The diastereomeric complexes **2**·*ent*-**1a** and **2**·**1a** have different distances between the substrate and the sensitizer (top, valence bond structures with key distances in red; bottom, three-dimensional structures based on DFT calculations). **b**, The association constants of complexes **2**·*ent*-**1a** and **2**·**1a** are different, as determined by nuclear magnetic resonance titration experiments in benzene-d_6 solution (blue, **2**·*ent*-**1a**; light blue, **2**·**1a**). **c**, The conversion of *ent*-**1a** to **1a**—that is, the inversion of compound *ent*-**1a**—occurs upon sensitization with **2** within 40 min and with a quantum yield of $\Phi = 0.52 \pm 0.03$.

in complex **2**·**1a**. Association constants for complexes were determined in benzene-d_6 (C_6D_6) solution by nuclear magnetic resonance titration[13] (Fig. 4b). In line with the DFT calculations, it was shown that the binding of compound *ent*-**1a** to the sensitizer is significantly stronger (association constant $K_a = 84 \pm 6\ M^{-1}$) than that of its enantiomer **1a** ($K_a = 18 \pm 2\ M^{-1}$).

Based on these data, we propose that the observed deracemization is due to the different overall sensitization rates of enantiomers *ent*-**1a** and **1a** as a consequence of their different association constants to the catalyst (ground-state thermodynamics) and different sensitization efficiencies (excited-state kinetics), which cooperate favourably. Triplet energy transfer is strongly distance-dependent[14,15], and in complex **2**·*ent*-**1a** the small distance between sensitizer and the substrate is responsible for a fast and quantitative triplet energy transfer. Sensitization of compound **1a** by an intramolecular mechanism, however, is disfavoured owing to its lower association to the sensitizer and to the increased distance between energy donor and acceptor compared to *ent*-**1a**. To substantiate this hypothesis, we determined the quantum yield of the inversion of compound *ent*-**1a** to compound **1a** under the optimized conditions in acetonitrile. Within less than 40 min, the compound was completely inverted. In a preparative run, enantiomer **1a** (96% e.e.) was obtained from *ent*-**1a** (99% e.e.) at 82% yield (see Supplementary Information, page 30). The quantum yield was high ($\Phi = 0.52 \pm 0.03$) and can be considered quantitative if one assumes[9] that upon excitation the planar allene triplet intermediate decays with a 50:50 ratio to **1a** or *ent*-**1a** (Fig. 4c). In other words, compound **1a** is formed as a major enantiomer in the deracemization with sensitizer **2** because its sensitization is highly inefficient, which in turn suggests that the intermolecular energy transfer is also slow, presumably owing to its minimally exothermic or even slightly endothermic nature[16]. Indeed, the racemization of compound *ent*-**1a** was catalysed by an achiral thioxanthone with a much lower quantum yield (see Supplementary Information, page 75) than that of the inversion, which was catalysed by **2**.

In summary, we have discovered an unprecedented catalytic enantioselective photochemical transformation[17–19] that enables deracemization of chiral compounds in a light-driven process. Minimal quantities of catalyst lead to a high multiplication of enantioselectivity, which is relevant for many fundamental scientific questions, including the origin of homochirality in biological systems[20]. The structural simplicity of the transformation is remarkable and should allow further expansion to other substrates. Representative substrate classes that have already been shown to be amenable to deracemization—albeit with low enantioselectivity—include chiral cyclopropanes[21,22] and chiral sulfoxides[23].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0755-1.

Received: 27 July 2018; Accepted: 24 October 2018;
Published online 12 December 2018.

1. Carreira, E. M. & Yamamoto, H. (eds) *Comprehensive Chirality* (Academic Press, Amsterdam, 2012).
2. Lorenz, H. & Seidel-Morgenstern, A. Processes to separate enantiomers. *Angew. Chem.* **53**, 1218–1250 (2014).
3. Palmans, A. R. A. Deracemisations under kinetic and thermodynamic control. *Mol. Syst. Des. Eng.* **2**, 34–46 (2017).
4. Rachwalski, M., Vermue, N. & Rutjes, F. P. J. T. Recent advances in enzymatic and chemical deracemisation of racemic compounds. *Chem. Soc. Rev.* **42**, 9268–9282 (2013).
5. Inoue, Y., Yokoyama, T., Yamasaki, N. & Tai, A. An optical yield that increases with temperature in a photochemically induced enantiomeric isomerization. *Nature* **341**, 225–226 (1989); erratum **342**, 958 (1989).
6. Maeda, R. et al. Planar-to-planar chirality transfer in the excited state. Enantiodifferentiating photoisomerization of cyclooctenes sensitized by planar-chiral paracyclophane. *J. Am. Chem. Soc.* **133**, 10379–10381 (2011).
7. Rodriguez, O. & Morrison, H. Photosensitized racemization of an optically active allene. *J. Chem. Soc. D* 679 (1971).
8. Bucher, G., Mahajan, A. A. & Schmittel, M. The photochemical C^2–C^6 cyclization of enyne–allenes: interception of the fulvene diradical with a radical clock ring opening. *J. Org. Chem.* **74**, 5850–5860 (2009).
9. Drucker, C. S., Toscano, V. G. & Weiss, R. G. A general method for the determination of steric effects during collisional energy transfer. Partial photoresolution of penta-2,3-diene. *J. Am. Chem. Soc.* **95**, 6482–6484 (1973).
10. Alonso, R. & Bach, T. A chiral thioxanthone as organocatalyst for enantioselective [2+2] photocycloaddition reactions induced by visible light. *Angew. Chem.* **53**, 4368–4371 (2014).
11. Gotthardt, H. & Hammond, G. S. Zur Photochemie des Allens. *Chem. Ber.* **108**, 657–663 (1975).
12. Shepard, M. S. & Carreira, E. M. Asymmetric photocycloadditions with an optically active allenylsilane: trimethylsilyl as a removable stereocontrolling group for the enantioselective synthesis of *exo*-methylenecyclobutanes. *J. Am. Chem. Soc.* **119**, 2597–2605 (1997).
13. Fielding, L. Determination of association constants (K_a) from solution NMR data. *Tetrahedron* **56**, 6151–6170 (2000).
14. Dexter, D. L. A theory of sensitized luminescence in solids. *J. Chem. Phys.* **21**, 836–850 (1953).
15. Turro, N. J., Ramamurthy, V. & Scaiano, J. C. *Modern Molecular Photochemistry of Organic Compounds* 411–413 (University Science Books, Sausalito, 2010).
16. Brennan, C. M., Caldwell, R. A., Elbert, J. E. & Unett, D. J. Nonvertical triplet excitation transfer to arylalkene acceptors: further evidence that double bond torsion is unimportant. *J. Am. Chem. Soc.* **116**, 3460–3464 (1994).
17. Silvi, M. & Melchiorre, P. Enhancing the potential of enantioselective organocatalysis with light. *Nature* **554**, 41–49 (2018).
18. Stephenson, C. R. J., Yoon, T. P. & MacMillan, D. W. C. (eds) *Visible Light Photocatalysis in Organic Chemistry* (Wiley-VCH, Weinheim, 2018).
19. Brimioulle, R., Lenhart, D., Maturi, M. M. & Bach, T. Enantioselective catalysis of photochemical reactions. *Angew. Chem.* **54**, 3872–3890 (2015).
20. Meierhenrich, U. *Amino Acids and the Asymmetry of Life* (Springer, Berlin, 2008)
21. Hammond, G. S. & Cole, R. S. Asymmetric induction during energy transfer. *J. Am. Chem. Soc.* **87**, 3256–3257 (1965).
22. Ouannès, C., Beugelmans, R. & Roussi, G. Asymmetric induction during transfer of triplet energy. *J. Am. Chem. Soc.* **95**, 8472–8474 (1973).
23. Balavoine, G., Jugé, S. & Kagan, H. B. Photoactivation optique du methyl p-tolyl sulfoxyde racemique par emploi d'un sensibilisateur chiral. *Tetrahedr. Lett.* **14**, 4159–4162 (1973).

© 2018 Springer Nature Limited. All rights reserved.

Acknowledgements Financial support by the Deutsche Forschungsgemeinschaft (DFG) via grant Ba1372/20 and GRK 1626 (T.B.) and through the Cluster of Excellence (EXC 1069) RESOLV project (S.M.H.) is acknowledged. A.H.-H. thanks the research training group (Graduiertenkolleg GRK) 1626 'Chemical Photocatalysis' for a scholarship. A.V.S. acknowledges the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for a post-doctoral research fellowship (BEX number 10744/13-4). C.B. is grateful to S. Grimme for discussions and for admission to computing clusters. We thank O. Ackermann and J. Kudermann for help with the high-performance liquid chromatography and gas–liquid chromatography analyses, A. Strauch for help with the circular dichroism measurements and A. Tröster for compounds for experimental comparison.

Reviewer information *Nature* thanks Y. Inoue and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.H.-H., A.B. and A.V.S. performed and analysed the experiments. A.H.-H., A.B., A.V.S. and T.B. designed the experiments. A.H.-H., A.B. and T.B. prepared the manuscript. S.M.H. performed all DFT calculations related to complexes **2**·*ent*-**1a** and **2**·**1a**. C.B. computed the theoretical chiroptical data for compound **1a**.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0755-1.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to T.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Here we describe the preparative procedure for the photoderacemization; a summary of the general conditions is included in Supplementary Information (pages 2–11), together with further experimental details (Supplementary Information, pages 12–29).

Procedure for the deracemization of compound 1a. Allene *rac*-**1a** (201 mg, 1.12 mmol, 1.00 equivalent) and (−)-thioxanthone catalyst **2** (12.1 mg, 28.1 μmol, 2.50 mol%) were dissolved in 112 ml of dry acetonitrile ($c = 10$ mM). The solution was degassed for 15 min by purging with argon under continuous sonication. The irradiation was performed in a previously described photoreactor[24] by irradiation with 16 fluorescent lamps that have an emission maximum at $\lambda = 420$ nm. The emission spectrum of the lamps can be found in Supplementary Information (page 3). After four hours, the irradiation was stopped and the solvent was removed under reduced pressure. The crude material was purified by column chromatography (silica, EtOAc). Product **1a** (199 mg, 99%, 95% e.e., 1.11 mmol) was obtained as a colourless solid.

Data availability

The findings of this study are available within the paper and its Supplementary Information.

24. Poplata, S. & Bach, T. Enantioselective intermolecular [2+2] photocycloaddition reaction of cyclic enones and its application in a synthesis of (−)-grandisol. *J. Am. Chem. Soc.* **140**, 3228–3231 (2018).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0700-3

Deconstructive diversification of cyclic amines

Jose B. Roque[1,2], Yusuke Kuroda[1,2], Lucas T. Göttemann[1] & Richmond Sarpong[1]*

Deconstructive functionalization involves carbon–carbon (C–C) bond cleavage followed by bond construction on one or more of the constituent carbons. For example, ozonolysis[1] and olefin metathesis[2,3] have allowed each carbon in C=C double bonds to be viewed as a functional group. Despite the substantial advances in deconstructive functionalization involving the scission of C=C double bonds, there are very few methods that achieve C(sp^3)–C(sp^3) single-bond cleavage and functionalization, especially in relatively unstrained cyclic systems. Here we report a deconstructive strategy to transform saturated nitrogen heterocycles such as piperidines and pyrrolidines, which are important moieties in bioactive molecules, into halogen-containing acyclic amine derivatives through sequential C(sp^3)–N and C(sp^3)–C(sp^3) single-bond cleavage followed by C(sp^3)–halogen bond formation. The resulting acyclic haloamines are versatile intermediates that can be transformed into various structural motifs through substitution reactions. In this way we achieve the skeletal remodelling of cyclic amines, an example of scaffold hopping. We demonstrate this deconstructive strategy by the late-stage diversification of proline-containing peptides.

The development of technologies that enable the late-stage diversification of bioactive, heterocycle-containing molecules (Fig. 1a) should facilitate access to underexplored chemical space[4]. Over the past two decades, considerable effort has been dedicated to the development of methods to functionalize C–H bonds at a late stage; this has enabled the fine-tuning of substituents on nitrogen heterocycles and enhanced their functional-group diversity[5,6] (Fig. 1b). In the medicinal chemistry community, there is a growing demand for methods that not only modify the periphery of molecules via C–H functionalization, but also modify their core framework in order to achieve skeletal diversity, a concept referred to as 'scaffold hopping'[7,8]. However, there are only a few methods known to achieve deconstructive functionalization, such as those involving unstrained cyclic amines[9–12]. One recent example generates an aldehyde intermediate that can be further transformed to install C–O, C–C and C–N bonds[13].

In this context, ring-opening chlorination or bromination would generate versatile intermediates en route to diverse cyclic amines by coupling to a variety of nucleophiles (Fig. 1b). Furthermore, the deconstructive halogenation of proline-containing peptides would furnish versatile intermediates for the late-stage diversification of these medicinally important entities[14]. Although ring-opening chlorination of cyclic amines is known, the existing methods to effect this transformation are limited to 3–5-membered, *N*-alkyl-substituted, cyclic amines because of competing *N*-dealkylation[15]. Recently, our laboratory introduced a silver-mediated deconstructive strategy to transform cyclic amine derivatives into fluorine-containing acyclic amine derivatives using Selectfluor, via the homolytic ring-opening of hemiaminal intermediates[16]. On the basis of mechanistic insights gained from our deconstructive fluorination protocol, we questioned whether it would be possible to access acyclic chloro- and bromoamines from cyclic amines using our deconstructive strategy. Upon examination of existing reports on silver-catalysed halogenation reactions, we recognized that the simple replacement of Selectfluor with *N*-halo reagents such as *N*-chlorosuccinimide (NCS) or *N*-bromosuccinimide (NBS) would be unproductive, presumably owing to their lower oxidation potential[17]. Therefore, a distinct approach would be required to oxidize Ag(I) to Ag(II) in order to achieve deconstructive bromination and chlorination.

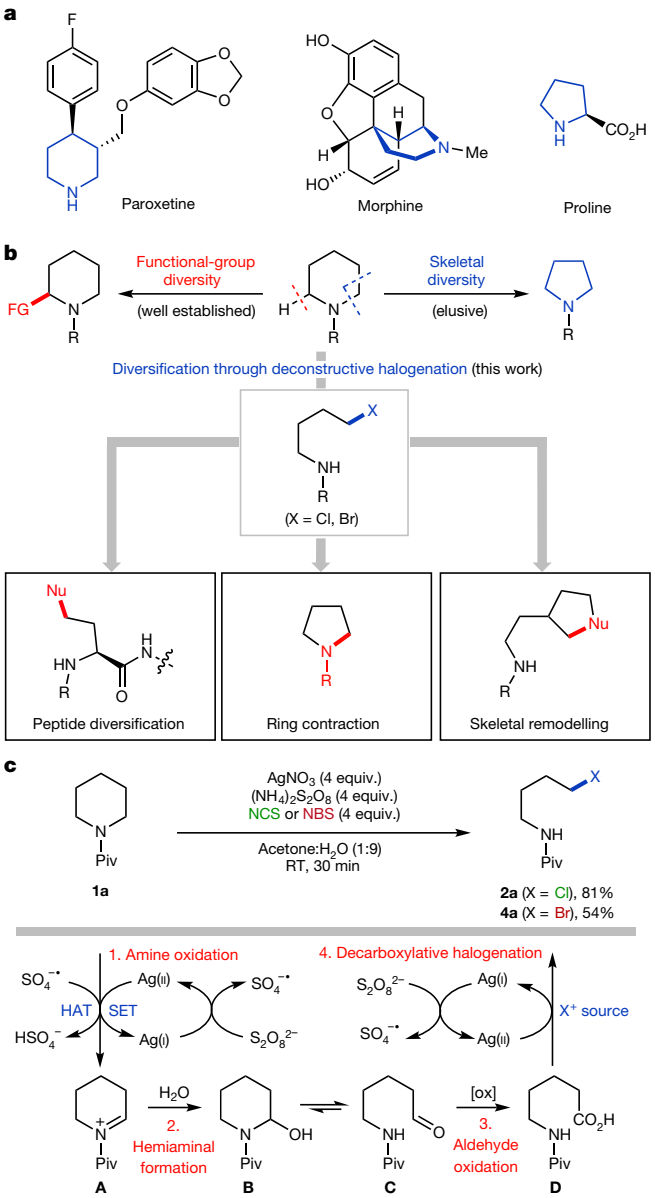


Fig. 1 | Development of a deconstructive halogenation of cyclic amines. **a**, Representative bioactive molecules containing saturated nitrogen heterocycles. **b**, Deconstructive halogenation enables diversification of saturated nitrogen heterocycles. **c**, Proposed mechanism for silver-mediated deconstructive halogenation. FG, functional group; HAT, hydrogen-atom transfer; Nu, nucleophile; [ox], oxidation; Piv, pivaloyl; SET, single-electron transfer.

[1]Department of Chemistry, University of California, Berkeley, CA, USA. [2]These authors contributed equally: Jose B. Roque, Yusuke Kuroda. *e-mail: rsarpong@berkeley.edu

© 2018 Springer Nature Limited. All rights reserved.

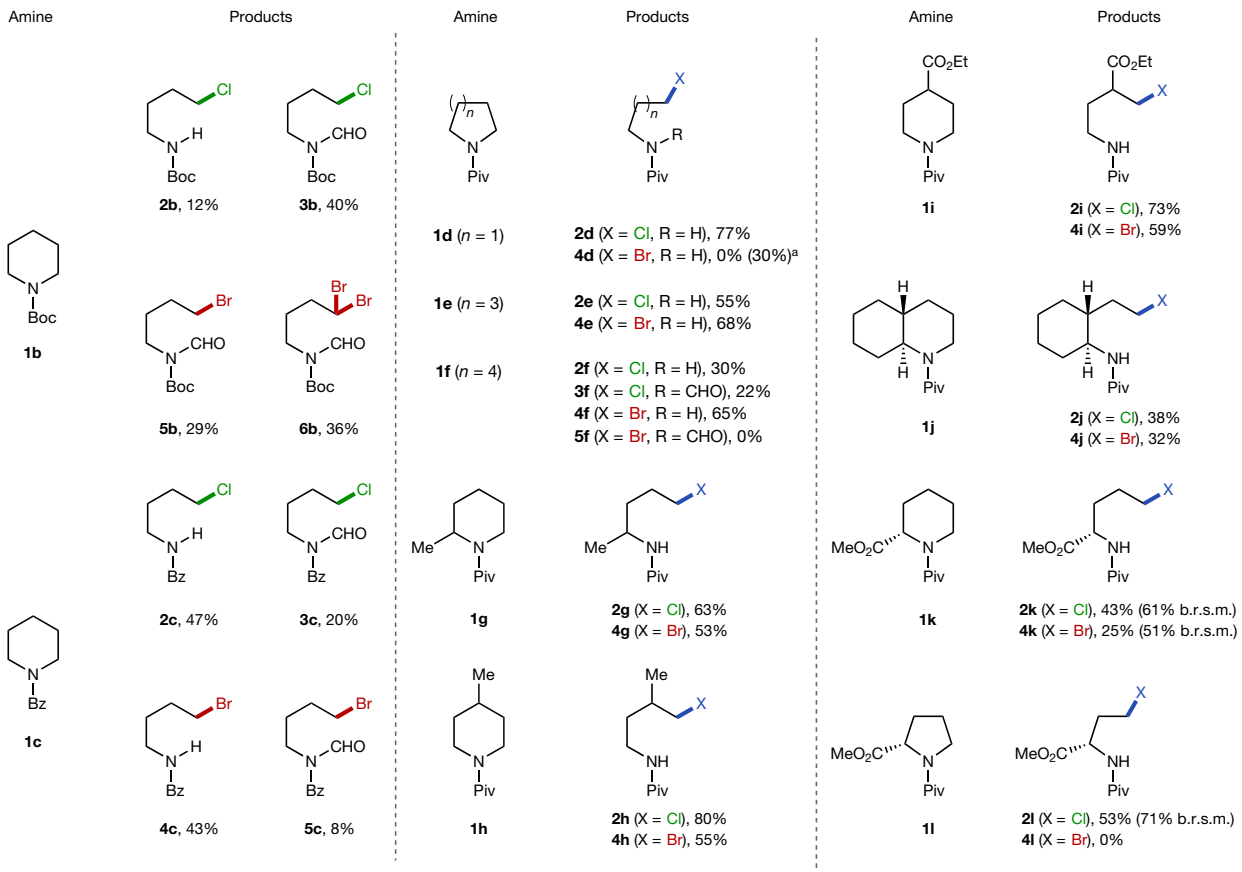


Fig. 2 | Scope of the cyclic amine in the deconstructive halogenation reaction. Only isolated yields are shown. Reactions were performed on a 0.2-mmol scale (based on **1**) under the conditions listed in Fig. 1c. b.r.s.m., based on recovered starting material. [a]5,6-dihydro-4*H*-1,3-oxazine was obtained (see Supplementary Information for details).

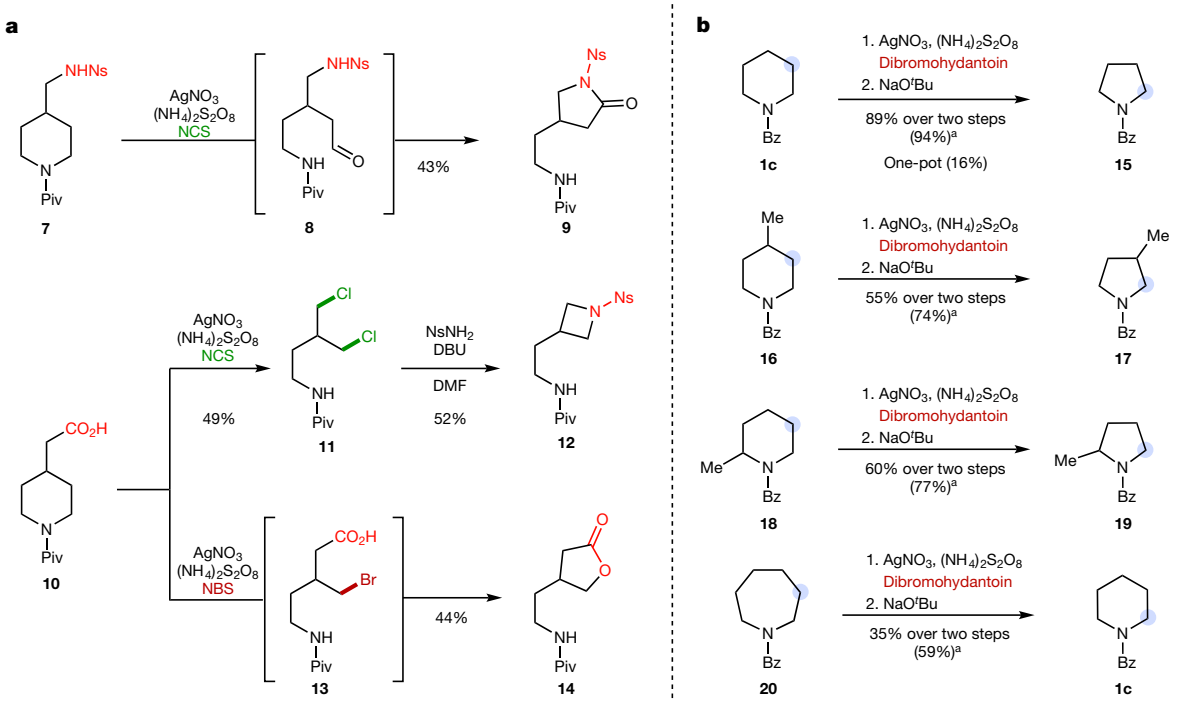


Fig. 3 | Applications of deconstructive halogenation. a, Skeletal remodelling of cyclic amines. **b**, Dehomologation of cyclic amines. [a]Yields in bracket represent the average yield per step.

© 2018 Springer Nature Limited. All rights reserved.

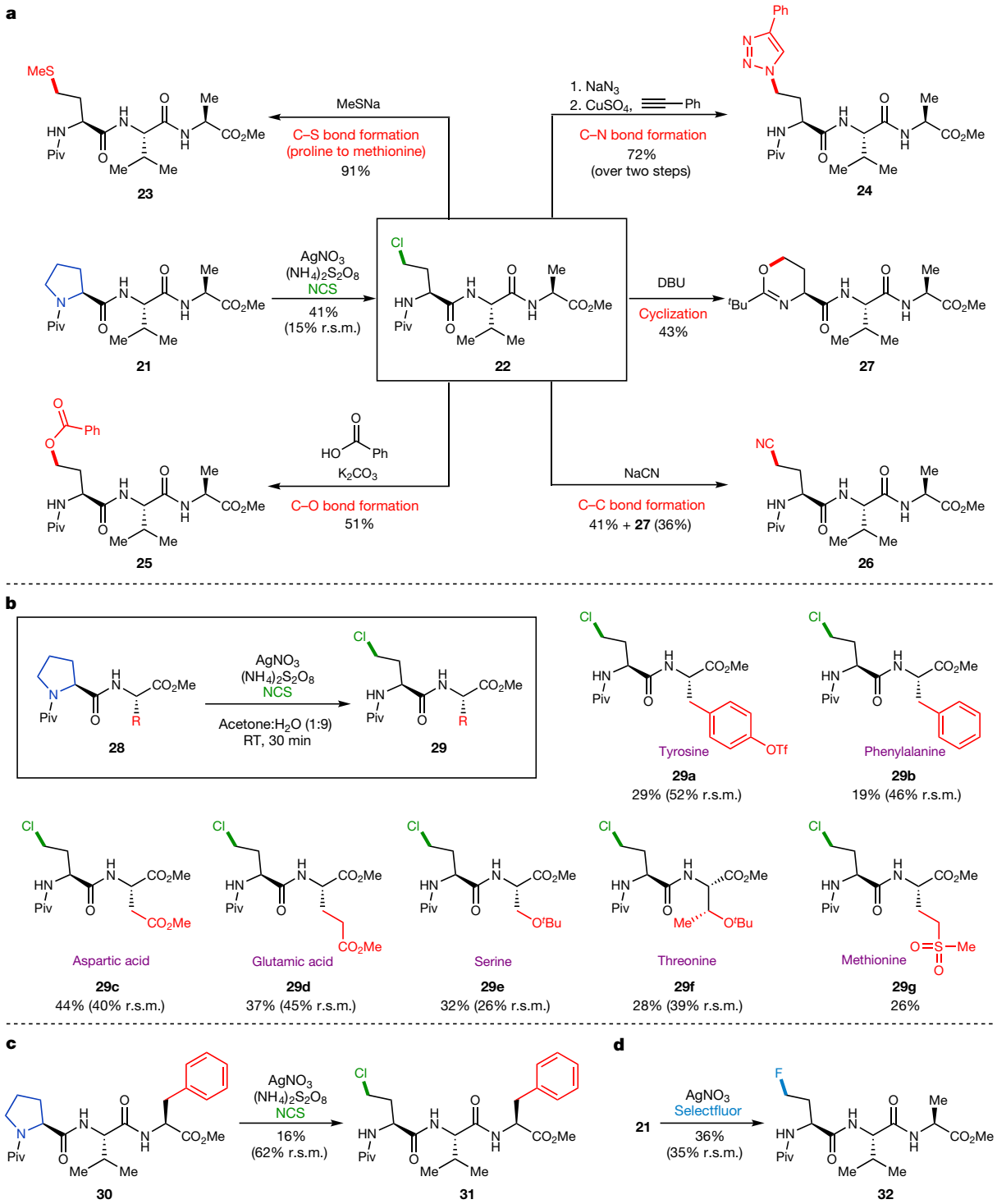


Fig. 4 | Deconstructive chlorination of l-proline-containing peptides. **a**, Deconstructive diversification of tripeptide **21**. **b**, The tolerance for oxidizable amino acid residues. **c**, Deconstructive chlorination of L-phenylalanine-containing tripeptide **30**. **d**, Deconstructive fluorination of tripeptide **21**. r.s.m., recovered starting material; Tf, trifluoromethanesulfonyl.

A detailed mechanistic proposal for our reaction sequence is depicted in Fig. 1c. We theorized that, consistent with existing precedent[18], Ag(I) will be oxidized to Ag(II) in the presence of persulfate anion, with concomitant disproportionation of the persulfate anion into a sulfate dianion and a sulfate radical anion. *N*-acylated cyclic amines **1** would then undergo a hydrogen-atom transfer with the resulting sulfate radical anion to give an α-amino radical[19]. Subsequent oxidation by Ag(II) via single-electron transfer would lead to iminium ion **A**. An alternative pathway is also possible, in which an Ag(II) species ($E°$ (Ag^{2+}/Ag^{+}) = +1.98 V vs saturated calomel electrode (SCE))[20] oxidizes *N*-acylated cyclic amines (for example, **1a**: (E_{pa} = +2.02 V vs SCE)) (Supplementary Fig. 1) to the radical cation via single-electron transfer, followed by hydrogen-atom transfer using the sulfate radical anion to generate the same iminium ion, **A**. The resulting iminium ion **A** would then be trapped by H_2O to give hemiaminal **B**. The heterolytic cleavage of the C–N bond would then occur through an equilibrium between hemiaminal **B** and aldehyde **C**, the latter being subsequently oxidized to carboxylic acid **D**[21], setting the stage for a silver-catalysed

© 2018 Springer Nature Limited. All rights reserved.

decarboxylative halogenation[17,22]. This strategy would represent a general method for deconstructive diversification, as the electrophile is independent of the initial redox cycle.

We commenced our investigations of the proposed deconstructive halogenation by evaluating a broad range of silver salts, halogenating reagents and solvent combinations. After extensive screening we identified the optimized conditions shown in Fig. 1c, using cheap and commercially available silver nitrate, ammonium persulfate and NCS in a 1:9 (v/v) mixture of acetone:H_2O at room temperature. Upon subjecting *N*-pivaloylpiperidine (**1a**) to the optimized conditions, we obtained 81% yield of the desired acyclic chlorinated product **2a**. Similarly, a bromine atom could be readily incorporated to afford **4a** in 54% yield by changing the electrophilic halogenating reagent to NBS. It is worth noting that this method can be performed without the strict exclusion of air. Control experiments established the importance of both silver and persulfate, as no formation of the desired chlorinated product was observed in the absence of the silver salt or the persulfate additive. The optimized conditions use four equivalents of silver nitrate, with lower quantities leading to reduced yields—presumably owing to substrate and/or product inhibition by binding to the silver salt (Supplementary Table 1).

With the optimized conditions in hand, we proceeded to investigate the scope of the deconstructive halogenation process (Fig. 2). An *N*-substituted piperidine derivative bearing a *tert*-butoxycarbonyl group (Boc, **1b**) gave the desired chlorinated products in a combined 52% yield of **2b** along with formimide product **3b**, which results from homolytic C–C bond cleavage of hemiaminal **B**[16]. Unlike the bulky pivaloyl group, which favours linear aldehyde **C** over hemiaminal **B** in the equilibration of the two species, the less sterically congested Boc group presumably favours **B** (Fig. 1c). Bromination using NBS led to a mixture of mono and dibrominated products **5b** and **6b** in 65% combined yield. Upon switching the group on nitrogen to benzoyl (Bz, **1c**), secondary amide products **2c** and **4c** were obtained as the major products along with formimide products **3c** and **5c**. In all cases, the secondary amide product and corresponding formimide are easily separated. Saturated heterocycles with various ring sizes (**1d**–**1f**) underwent deconstructive halogenation in moderate to good yields (55%–77% combined yield), although the deconstructive bromination of **1d** led to 5,6-dihydro-4*H*-1,3-oxazine through autocyclization of desired alkyl bromide **4d** (Supplementary Information)[23]. Substituents at the 2- and 4-position on piperidines are also well tolerated (**1g**–**1i**, 53%—80%). Polycyclic compounds such as **1j** are also readily functionalized, paving the way for late-stage derivatization in more complex polycyclic frameworks. Halogenated amino acid derivatives (**2k**, **2l** and **4k**) are accessed in three steps from L-proline and L-pipecolic acid, which may serve as versatile intermediates to other unnatural amino acids.

Next, the skeletal remodelling of piperidine scaffolds bearing other reactive groups was examined (Fig. 3a). Oxidative ring-opening of **7** followed by the coupling of the pendant 2-nitrobenzenesulfonamide (NsNH) nucleophile with the incipient aldehyde group in **8** ultimately yielded the corresponding lactam **9**. The choice of halogenating reagent led to divergence in the products that were formed. For example, when carboxylic acid **10** was subjected to the deconstructive chlorination conditions, dichloro compound **11** was obtained through decarboxylative[17] and deconstructive chlorination, and was directly transformed to azetidine **12** via double nucleophilic displacement with $NsNH_2$. Alternatively, when NBS was used as the halogenating agent, in-situ-generated alkyl bromide **13** reacted with the carboxylic acid group to form the corresponding lactone **14** in 44% yield.

Given the aforementioned importance of scaffold hopping in cyclic systems[7,8], we have also pursued the ring contraction of piperidines (Fig. 3b). There are few reports that detail the ring contraction of piperidines to pyrrolidines[24–26]. Deconstructive bromination of *N*-benzoyl piperidine (**1c**) with dibromohydantoin followed by cyclization of the resulting bromoamine with sodium *tert*-butoxide furnished *N*-benzoyl pyrrolidine (**15**) in 89% yield in just two steps (94% average yield per step), with only one chromatographic purification step. Notably, this process can also be conducted in one pot, albeit in lower yield (unoptimized) owing to the competing displacement of the newly installed halogen group by the imide byproduct from the halogenating reagent. This ring-contraction process also proceeds for a series of simple cyclic amines, such as 2- and 4-methyl substituted piperidines and azepane (**16**, **18** and **20**, 35%–60% yield over two steps). These results demonstrate a powerfully direct approach to achieving deep-seated structural modifications.

The virtue of this methodology is evident in the deconstructive functionalization and diversification of peptides[27]. As shown in Fig. 4a, L-proline-containing tripeptide **21** underwent ring-opening chlorination in 41% yield, along with 15% of recovered starting material. Chlorinated peptide **22** is easily transformed into a variety of products. For example, treatment of **22** with sodium methylthiolate afforded **23** in 91% yield, constituting the conversion of a proline residue into the corresponding methionine residue in only two steps. Alternatively, C–N bond formation can be achieved by the treatment of **22** with sodium azide, and in this way a proline residue or polypeptides that bear a cyclic amine (for example, L-pipecolic acid) can be converted into a site for azide-based biorthogonal click chemistry[28]. In a demonstration of this strategy, **22** was azidated and then subjected to copper-catalysed azide–alkyne cycloaddition to afford triazole **24** in 72% yield over the two steps. In addition, C–O bond formation is also easily achieved by displacement of the halogen group with benzoic acid. Treatment of **22** with NaCN in dimethylformamide led to nitrile **26** as the major product along with 5,6-dihydro-4*H*-1,3-oxazine **27** in 36% yield, demonstrating the feasibility of C–C bond formation. Cyclized product **27** is obtained as the sole product when **22** is treated with 1,8-diazabicyclo (5.4.0)undec-7-ene (DBU).

Additionally, we evaluated the functional-group tolerance of the deconstructive chlorination process. As shown in Fig. 4b, a variety of dipeptides bearing potentially oxidizable amino acid residues participate in this deconstructive protocol (**29a**–**29f**, 19%–44%). It is worth noting that the proline residue can be preferentially oxidized over the benzylic position (**29a**, **29b**) and C–H bonds of the activated aliphatic side chains bearing oxygen heteroatoms (**29e**, **29f**). A dipeptide bearing a methionine residue underwent deconstructive chlorination with oxidation of the thioether to the corresponding sulfone (**29g**). Therefore, like many other oxidative processes[29,30], deconstructive halogenation leads to a competing reaction with the sulfur group of methionine. Additionally, deconstructive chlorination of the challenging tripeptide substrate **30** proceeded to furnish ring-opened product **31** in a 16% yield, along with 62% of recovered starting material (Fig. 4c). Given that the current methodology involves a mechanistic change compared with our previous deconstructive fluorination strategy[16]—namely the incorporation of a heterolytic C–N cleavage (**B**→**C**, Fig. 1c)—the over-oxidation of the hemiaminal intermediate **B** is generally avoided, as evidenced by the ring-opening fluorination of **21** to give fluorinated tripeptide **32** using this new method (Fig. 4d). Despite the lower yields obtained in the presence of these reactive residues, the deconstructive protocol provides an expedient approach to a novel range of peptides without the need for their de novo synthesis.

Saturated heterocycles remain a prevalent structural motif that is found in a large percentage of bioactive organic molecules such as pharmaceuticals. We anticipate that deconstructive functionalization strategies will provide access to wide-ranging structural diversity at a late stage in the preparation of bioactive molecules.

Data availability

All data supporting the findings of this study are available within the paper and its Supplementary Information, or from the corresponding author upon reasonable request.

Received: 21 May 2018; Accepted: 23 October 2018;
Published online 31 October 2018.

© 2018 Springer Nature Limited. All rights reserved.

1. Bailey, P. S. The reactions of ozone with organic compounds. *Chem. Rev.* **58**, 925–1010 (1958).
2. Hoveyda, A. H. & Zhugralin, A. R. The remarkable metal-catalysed olefin metathesis reaction. *Nature* **450**, 243–251 (2007).
3. Vougioukalakis, G. C. & Grubbs, R. H. Ruthenium-based heterocyclic carbene-coordinated olefin metathesis catalysts. *Chem. Rev.* **110**, 1746–1787 (2010).
4. Galloway, W. R. J. D., Isidro-Llobet, A. & Spring, D. R. Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat. Commun.* **1**, 80 (2010).
5. Chen, W., Ma, L., Paul, A. & Seidel, D. Direct α-C–H bond functionalization of unprotected cyclic amines. *Nat. Chem.* **10**, 165–169 (2018).
6. Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016).
7. Sun, H., Tawa, G. & Wallqvist, A. Classification of scaffold-hopping approaches. *Drug Discov. Today* **17**, 310–324 (2012).
8. Hu, Y., Stumpfe, D. & Bajorath, J. Recent advances in scaffold hopping. *J. Med. Chem.* **60**, 1238–1246 (2017).
9. Shawcross, A. P. & Stanforth, S. P. Reaction of *N*-nitroaryl-1,2,3,4-tetrahydroisoquinoline derivatives with oxygen. *J. Heterocycl. Chem.* **27**, 367–369 (1990).
10. Han, G., McIntosh, M. C. & Weinreb, S. M. A convenient synthetic method for amide oxidation. *Tetrahedr. Lett.* **35**, 5813–5816 (1994).
11. Ito, R., Umezawa, N. & Higuchi, T. Unique oxidation reaction of amides with pyridine-*N*-oxide catalyzed by ruthenium porphyrin: direct oxidative conversion of *N*-acyl-l-proline to *N*-acyl-l-glutamate. *J. Am. Chem. Soc.* **127**, 834–835 (2005).
12. Kaname, M., Yoshifuji, S. & Sashida, H. Ruthenium tetroxide oxidation of cyclic *N*-acylamines by a single layer method: formation of ω-amino acids. *Tetrahedr. Lett.* **49**, 2786–2788 (2008).
13. Osberger, T. J., Rogness, D. C., Kohrt, J. T., Stepan, A. F. & White, M. C. Oxidative diversification of amino acids and peptides by small-molecule iron catalysis. *Nature* **537**, 214–219 (2016).
14. Henninot, A., Collins, J. C. & Nuss, J. M. The current state of peptide drug discovery: Back to the future? *J. Med. Chem.* **61**, 1382–1414 (2018).
15. Yu, C. et al. Selective ring-opening of *N*-alkyl pyrrolidines with chloroformates to 4-chlorobutyl carbamates. *J. Org. Chem.* **82**, 6615–6620 (2017).
16. Roque, J. B., Kuroda, Y., Göttemann, L. T. & Sarpong, R. Deconstructive fluorination of cyclic amines by carbon–carbon cleavage. *Science* **361**, 171–174 (2018).
17. Wang, Z. et al. Silver-catalyzed decarboxylative chlorination of aliphatic carboxylic acids. *J. Am. Chem. Soc.* **134**, 4258–4263 (2012).
18. Anderson, J. M. & Kochi, J. K. Silver(i)-catalyzed oxidative decarboxylation of acids by peroxydisulfate. Role of silver(ii). *J. Am. Chem. Soc.* **92**, 1651–1659 (1970).
19. Dai, C., Meschini, F., Narayanam, J. M. R. & Stephenson, C. R. J. Friedel–Crafts amidoalkylation via thermolysis and oxidative photocatalysis. *J. Org. Chem.* **77**, 4425–4431 (2012).
20. Po, H. N. Heterocyclic and macrocyclic amine complexes of silver(ii) and silver(iii). *Coord. Chem. Rev.* **20**, 171–195 (1976).
21. Edwards, J. O. & Gallopo, A. R. Kinetics and mechanisms of the spontaneous and metal-modified oxidations of ethanol by peroxydisulfate ion. *J. Org. Chem.* **36**, 4089–4096 (1971).
22. Tan, X. et al. Silver-catalyzed decarboxylative bromination of aliphatic carboxylic acids. *Org. Lett.* **19**, 1634–1637 (2017).
23. Reddy, D. N. & Prabhakaran, E. N. Synthesis and isolation of 5,6-dihydro-4*H*-1,3-oxazine hydrobromides by autocyclization of *N*-(3-bromopropyl)amides. *J. Org. Chem.* **76**, 680–683 (2011).
24. Bandarage, U. K. & Davies, R. J. A new synthesis of spiropyrrolidine–tetralones via an unexpected formal ring-contraction of 4-disubstituted piperidine to 3-disubstituted pyrrolidine. *Tetrahedr. Lett.* **51**, 6415–6417 (2010).
25. Aldmairi, A. H., Griffiths-Jones, C., Dupauw, A., Henderson, L. & Knight, D. W. Piperidines from acid-catalysed cyclisations: pitfalls, solutions and a new ring contraction to pyrrolidines. *Tetrahedr. Lett.* **58**, 3690–3694 (2017).
26. Sengupta, S. & Mehta, G. Late stage modification of peptides via C–H activation reactions. *Tetrahedr. Lett.* **58**, 1357–1372 (2017).
27. Wang, F., He, Y., Tian, M., Zhang, X. & Fan, X. Synthesis of α-formylated *N*-heterocycles and their 1,1-diacetates from inactivated cyclic amines involving an oxidative ring contraction. *Org. Lett.* **20**, 864–867 (2018).
28. Kolb, H. C., Finn, M. G. & Sharpless, K. B. Click chemistry: diverse chemical function from a few good reactions. *Angew. Chem. Int. Ed.* **40**, 2004–2021 (2001).
29. Shechter, Y., Burstein, Y. & Patchornik, A. Selective oxidation of methionine residues in proteins. *Biochemistry* **14**, 4497–4503 (1975).
30. Lin, S. et al. Redox-based reagents for chemoselective methionine bioconjugation. *Science* **355**, 597–602 (2017).

Acknowledgements This work was supported by the National Institutes of Health (NIH; NIGMS RO1 086374). J.B.R. thanks the NIH for a graduate diversity supplement fellowship (NIGMS RO1 086374). Y.K. thanks the Japan Society for the Promotion of Science (JSPS) for an Overseas Research Fellowship. L.T.G. thanks LMU PROSA and DAAD for financial support. We thank J. Derrick for assistance with electrochemical measurements.

Author contributions J.B.R. and Y.K. conceived the research and designed the experiments. J.B.R., Y.K. and L.T.G. performed the experiments. R.S. directed the project. J.B.R., Y.K. and R.S. wrote the manuscript.

Competing interests J.B.R., Y.K., L.T.G. and R.S. are listed as inventors on an initial patent application describing the silver-mediated deconstructive halogenation of cyclic amines and subsequent transformations (052103-515P01US).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0700-3.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0757-z

Assessing the efficiency of changes in land use for mitigating climate change

Timothy D. Searchinger[1,2]*, Stefan Wirsenius[3], Tim Beringer[4] & Patrice Dumas[5,6]

Land-use changes are critical for climate policy because native vegetation and soils store abundant carbon and their losses from agricultural expansion, together with emissions from agricultural production, contribute about 20 to 25 per cent of greenhouse gas emissions[1,2]. Most climate strategies require maintaining or increasing land-based carbon[3] while meeting food demands, which are expected to grow by more than 50 per cent by 2050[1,2,4]. A finite global land area implies that fulfilling these strategies requires increasing global land-use efficiency of both storing carbon and producing food. Yet measuring the efficiency of land-use changes from the perspective of greenhouse gas emissions is challenging, particularly when land outputs change, for example, from one food to another or from food to carbon storage in forests. Intuitively, if a hectare of land produces maize well and forest poorly, maize should be the more efficient use of land, and vice versa. However, quantifying this difference and the yields at which the balance changes requires a common metric that factors in different outputs, emissions from different agricultural inputs (such as fertilizer) and the different productive potentials of land due to physical factors such as rainfall or soils. Here we propose a carbon benefits index that measures how changes in the output types, output quantities and production processes of a hectare of land contribute to the global capacity to store carbon and to reduce total greenhouse gas emissions. This index does not evaluate biodiversity or other ecosystem values, which must be analysed separately. We apply the index to a range of land-use and consumption choices relevant to climate policy, such as reforesting pastures, biofuel production and diet changes. We find that these choices can have much greater implications for the climate than previously understood because standard methods for evaluating the effects of land use[4–11] on greenhouse gas emissions systematically underestimate the opportunity of land to store carbon if it is not used for agriculture.

We define a more 'carbon efficient' use of land as one that increases the capacity of global land to store carbon and reduce greenhouse gas emissions (GHGs) overall, while meeting the same global food demand. For example, producing more crops, meat or milk on one hectare of land increases this carbon efficiency by increasing the global capacity to spare forests and other habitats while producing the same quantity of food. Gains in efficiency increase capacity to generate valuable outputs but do not by themselves guarantee how the added capacity will be used—for example, for more carbon or more food—or how other people might react owing to market forces. Yet because land supply is fixed, only increasing its efficiency can allow the world to meet both climate and food goals.

Governments, companies and individuals are making land-use decisions at least partially directed at reducing GHGs. Questions include whether to encourage conversion of cropland to forest or bioenergy, what targets to set for national emissions from land use and how to reduce the carbon footprint of diets or food supply chains. Yet standard evaluation methods, as discussed below and in more detail in Supplementary Information, do not properly reflect the land's opportunity to store carbon if it is not used for agriculture, which we call its carbon storage opportunity cost. They can therefore encourage inefficient results that reduce the global capacity to store carbon.

For example, typical lifecycle assessments (LCAs), which estimate the GHG costs of a food's consumption, only estimate land-use demands in hectares without translating them into carbon costs[4,5]. Other LCAs consider land-use carbon costs only if a food is directly produced by clearing new land[6,7], or only for specific crops, meat or milk, where both that food and agricultural land overall are expanding[8–10]. Such approaches assign no land-use carbon costs to most of the world's food production because previously converted agricultural lands have no carbon storage opportunity cost[12] (Supplementary Information).

Physical optimization models[13,14] can estimate where agricultural expansion should occur to minimize carbon costs, by assuming likely crop yields of every hectare in a study area. Such models can count carbon storage opportunity costs, but they cannot account for the variability in carbon storage or crop yields in real hectares or estimate the effects of changes in their yields, output types or production methods (Supplementary Information).

Economic models provide a common approach to estimating how conversion of cropland to biofuels or forest affects carbon stored elsewhere, called 'leakage' or 'indirect land-use change' (ILUC). However, these models do not calculate the true efficiency of the changes to the hectare analysed (for example, reforesting cropland) because the models also factor in how resulting increases in food prices cause changes on other land, by other people and at others' expense. Such changes may include lower GHGs through reductions in global food consumption and, although disputed, through simulated increases in the yields (efficiencies) of other farmland[15]. Such estimated 'benefits', paid for by global consumers, result from the decline in food production on the hectare whose use was deliberately changed, not from its gain in forest or bioenergy, and would therefore occur even if that hectare became supremely inefficient by producing nothing at all.

To appreciate the distinction, we imagine a possible economic analysis of a strange climate policy banning all cars except petrol-guzzling, expensive, luxury SUVs (sport utility vehicles). The efficiency of driving would decline, increasing emissions per kilometre. However, if the cost of driving rose high enough, an economic model might estimate overall GHG savings by forcing many people to stay at home and others to switch to public transit. Even if these outcomes were real, these switches would not make SUVs more efficient than economy cars.

The actual efficiency of driving matters because governments can reduce GHGs more generally by using fuel taxes and transit subsidies to encourage less travel and higher use of mass transit while also requiring vehicles that are more fuel-efficient. Similarly, if governments wished to use higher prices to reduce food consumption and spur yield gains, they could reduce GHGs more using taxes and subsidies while encouraging only efficient land-use changes (LUCs). To implement such policies, however, governments need to know which LUCs are more efficient in themselves.

[1]Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ, USA. [2]World Resources Institute, Washington, DC, USA. [3]Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg, Sweden. [4]Integrative Research Institute on Transformations of Human Environment Systems (IRI THESys), Humboldt-Universität zu Berlin, Berlin, Germany. [5]Centre International de Recherche sur l'Environnement et le Développement (CIRED), Montpelier, France. [6]Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Montpelier, France. *e-mail: tsearchi@princeton.edu

© 2018 Springer Nature Limited. All rights reserved.

Table 1 | COCs and global PEMs of major crop and livestock products

	COC[a] (kg CO_2 per kg fresh weight)	PEMs (kg CO_2e per kg fresh weight)	Total (kg CO_2e per kg fresh weight)	Total (g CO_2e per kcal[c])	Total (kg CO_2e per kg protein)
Maize	2.1	0.46	2.6	0.82	29
Rice (rough)	2.6	2.17	4.8	2.0	69
Wheat	1.9	0.69	2.6	0.9	23
Cassava	1.7	0.04	1.7	1.6	160
Potato	0.6	0.09	0.7	1.1	38
Soybeans	5.9	0.26	6.1	1.5	17
Pulses	10.5	0.55	11	3.1	47
Vegetable oils	9.7	1.3	11	1.2	Not applicable
Beef[b]	144	44	188	102	1,250
Cow milk	6.2	2.3	8.4	13.1	260
Pork	14	5.5	20	9.4	150
Poultry meat	11	3.7	14	8.4	110

Values are calculated using the carbon loss method and 4% time discounting.
[a]Includes peatland emissions.
[b]Average, including meat from dairy animals.
[c]1 kcal = 4,184 J.

Our carbon benefits index provides such a measure, expressing benefits as kilograms of CO_2 equivalent (CO_2e) emissions per hectare (ha; 1 ha = 10^4 m^2) per year. The index first incorporates the outputs of a hectare that are directly quantifiable in carbon terms. These include any changes in carbon storage on site, as well as net reductions in GHGs from displacing fossil fuels with bioenergy.

The challenge arises in calculating the carbon benefits of producing foods, whose carbon is consumed. The index values them according to the emissions that are avoided from their production elsewhere. The core assumption is that if one hectare did not produce a food, it would be produced elsewhere at its global-average carbon costs. By holding consumption and other production systems fixed, the index calculates only changes in the efficiency of the hectare analysed.

We call the land cost of replacing each food its 'carbon opportunity cost' (COC), and calculate it using two methods. In the first method, the 'carbon loss' method, the COC is equal to the global carbon loss from plants and soils generated by producing each crop to date (the numerator), divided by the global production (the denominator), and is expressed as kilograms of CO_2e per kilogram of crop. For each meat or milk, the COC is equal to the sum of COCs of the feeds needed to produce it (including lost carbon on pasture for ruminants). The COCs of bioenergy feed by-products equal the COCs of the crops that they displace.

The second method is the 'carbon gain' method, in which we estimate the quantity of carbon that could be sequestered annually if the average productive capacity of land used to produce a kilogram of each food globally were instead devoted to regenerating forest. The carbon loss method is generally more appropriate in a world of expanding cropland, but the carbon gain method could apply where increasing yields could only increase carbon by rebuilding forests.

Because carbon losses of native vegetation occur quickly yet food production could continue indefinitely, we calculate a present discount value of both the numerator and denominator. The choice of rate to discount the costs and benefits of changes in the future is a question of climate policy. We use 4% in our central scenario, in part to match the implicit approach of US biofuel policies (Supplementary Information)[15].

Table 1 presents COCs calculated with the carbon loss method for a sample of products using fresh weights, with a fuller list of 64 products in Extended Data Tables 1, 2 (including COCs using dry matter). COCs from the carbon gain method are mostly similar using a 4% discount rate and do not alter the directional results of our examples (Supplementary Tables 5–9). The COCs reflect the different average yields and native carbon stocks of lands used by different crops. For example, soybean COCs are 2.8 times larger than maize COCs because soybean yields are lower, even though the crops use similar lands. However, wheat and maize COCs are similar despite 40% lower wheat yields because wheat is grown overall on land with lower native carbon stocks.

Our core replacement assumption implies that if the rate of a food's recurring production emissions (PEMs) on one hectare is lower or higher than the global average, the difference decreases or increases global PEMs. When calculating global-average PEMs for all foods (Extended Data Tables 1, 2), we factor that difference into the index. The index can therefore calculate the net GHG effects of altering yields by changing fertilizer or livestock feeds, which alter N_2O and CH_4 emissions.

In summary, the total carbon benefits of a hectare of land is equal to the sum of: (1) the opportunity that its food production provides to store carbon elsewhere (COC × yield), (2) its savings or increase in global PEMs, (3) its annual change in soil and plant carbon storage and (4) any net savings in fossil emissions due to bioenergy generated (see equation (1) in Methods). The efficiency of an LUC depends on the gain or loss in carbon benefits.

The index can also evaluate the carbon efficiency of consumption by assuming that production systems are fixed. One individual's change in consumption therefore alters global consumption and aggregate production by that amount. The cost of a food is equal to its COC plus its PEMs (see equation (2) in Methods).

The index separates the efficiency of consumption from the efficiency of each hectare's production into different analyses. The higher a product's COC, the costlier its consumption, but also the more beneficial its production. For example, consuming a kilogram of beef costs more carbon than consuming a kilogram of soybeans, but producing a kilogram of beef generates more benefits because it frees up more carbon storage capacity elsewhere, assuming fixed demand.

We supply a Carbon Benefits Calculator (provided as Supplementary Data) for users to evaluate the efficiency of changes in real hectares using site-specific information and changes to discount rates, COCs and other parameters. We also apply our index to production and consumption choices that are important to global climate policy in the following examples (see Supplementary Information for full sources of information used in the examples).

First, we consider production changes from Brazil grazing land. In Brazil, because of low yields of beef from extensive cattle grazing, proposals exist either to convert pastures to cropland for soybeans or to sugarcane for ethanol, or to intensify pasture management to help meet the expected increases of about 80% in global beef demand by 2050[1,16]. We consider which changes would produce more carbon benefits. Cardoso et al.[17] categorized beef production in the

© 2018 Springer Nature Limited. All rights reserved.

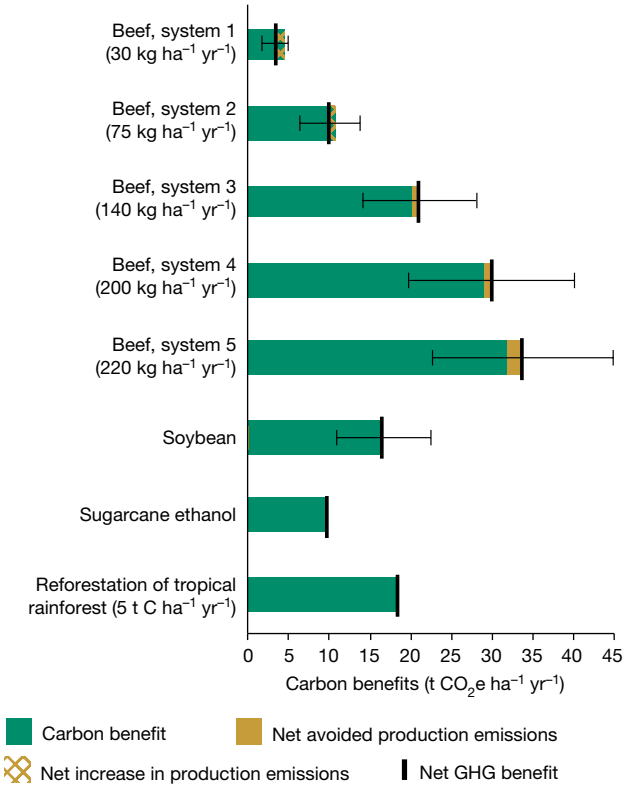


Fig. 1 | Carbon benefits of different potential uses of Brazilian grazing land. Error bars reflect the range of literature estimates of vegetation and soil carbon stocks used in part to derive the COCs.

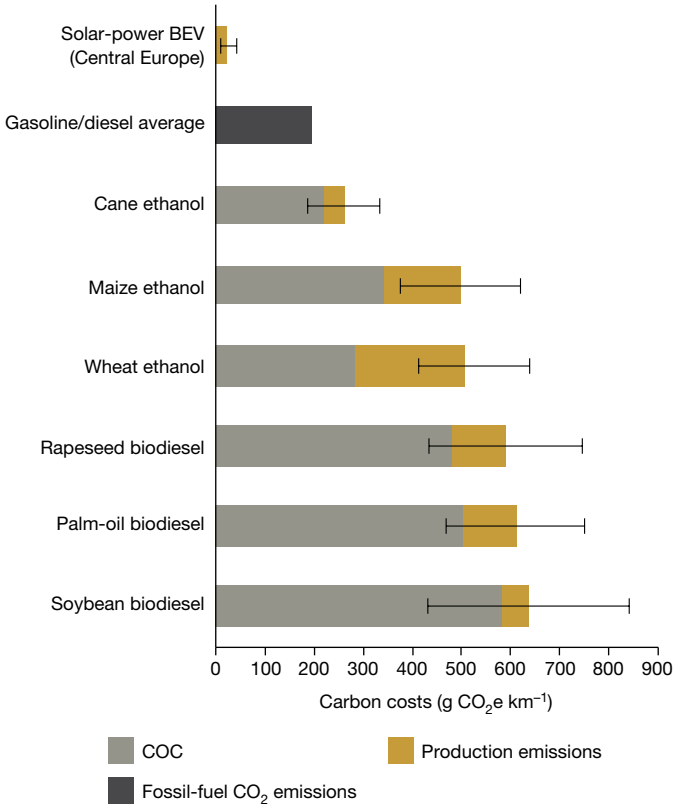


Fig. 2 | Carbon costs of different fuel sources (per kilometre driven) based on the carbon benefits index. Error bars reflect the range of literature estimates of vegetation and soil carbon stocks used in part to derive the COCs. BEV, battery electric vehicles.

Cerrado region in Brazil into five systems with increasing beef yields from 30 kg ha^{-1} yr^{-1} to 220 kg ha^{-1} yr^{-1} on the basis of grazing practices, healthcare, fertilization and replanting frequency, and uses of legumes or crop supplements. We find that for grazing land using system 1 (30 kg ha^{-1} yr^{-1}, carcass weight) shifting to sugarcane ethanol increases carbon benefits (Fig. 1). However, the more commonly used system 2 (75 kg ha^{-1} yr^{-1}) generates roughly the same benefits as system 1, whereas system 3 (150 kg ha^{-1} yr^{-1}) produces much greater carbon benefits[17]. Shifting to soybean production at average Brazilian yields would produce more benefits than grazing system 2, but less than system 3.

By contrast, reforesting pastures at 5 t C ha^{-1} yr^{-1} would increase carbon benefits by a factor of five in the Atlantic Coastal Rainforest region. Grazing system 1 is mostly used in this region at present and pastures are difficult to intensify because they are mainly located on steep terrain.

Factoring in the land's COCs, we find that shifting from system 1 to system 3 increases benefits six times, in contrast to the merely twofold gain[16] from PEMs only. Shifting from grazing system 2 to system 3 provides annual benefits equivalent to those of temperate forest growth (about 3 t C ha^{-1} yr^{-1})[18,19]. Extensive systems that use arid, native grasslands, including nomadic systems, can still be efficient despite producing little beef and few carbon benefits because they also sacrifice little opportunity to store carbon (Supplementary Information).

Second, we consider production changes related to intensification. By examining several plausible examples, we find that increasing crop inputs usually saves more GHGs through reduced land demands than the increase in GHGs because of higher PEMs. Examples include adding 75 kg ha^{-1} yr^{-1} of nitrogen to maize in West Africa using flooded, irrigated rice, rather than upland rice, despite its higher methane emission, and comparing conventional versus organic bean production in Sweden (Extended Data Fig. 1).

Third, we consider production changes related to biofuel production. Carbon benefits from cropland with a rotation of maize and soybeans at average Iowa yields (22 t CO$_2$e ha^{-1} yr^{-1}) greatly exceed those of ethanol production either from maize (9 t CO$_2$e ha^{-1} yr^{-1}) or from perennial grasses (17.5 t CO$_2$e ha^{-1} yr^{-1}) (assuming optimistic grass dry-matter yields[20] of 17 t ha^{-1} yr^{-1} and 0.6 t C ha^{-1} yr^{-1} soil carbon sequestration[21]; Extended Data Fig. 2). For maize ethanol, feed by-products provide two-thirds of the benefits. Perennial grasses for ethanol would have to achieve implausibly high dry-matter yields of 32 t ha^{-1} yr^{-1} to match the benefits of maize–soybean rotations.

Fourth, we consider consumption changes related to biofuels. We estimate that the total GHG costs of consuming biofuels, rather than gasoline or diesel, range from 35% more for sugarcane ethanol to 230% more for soybean biodiesel (Fig. 2). Using Central European solar power to run battery electric vehicles generates only 9% of the GHGs of sugarcane ethanol, mostly through battery production (Supplementary Information). Our biofuel COC estimates are equivalent to ILUC estimates if crops diverted to biofuels (after deducting by-products) are fully replaced at the average global carbon loss per kilogram of crop. Our estimates range from 100 g MJ^{-1} to 300 g MJ^{-1} of CO$_2$ emissions for biofuels from different feedstocks—higher than gasoline or diesel emissions, even without counting their PEMs (Extended Data Table 3). Our estimates are mostly 6–14 times higher than those of economic models commissioned by California and the European Commission (Supplementary Table 4).

Last, we consider consumption changes related to shifting diets. LCAs have long estimated GHG benefits from diet shifts away from ruminant products (box 8 in ref. [12]), but typically assign little or no GHG costs to land requirements[5,8,22]. By applying the carbon benefits index, we find global-average GHG costs of dairy and beef about 3–4 times higher than previous estimates by the UN Food and Agriculture Organization[8] (Supplementary Information), which only include land-use GHGs from each year's agricultural expansion. We estimate the total GHG costs of average Northern European diets[22] at more than 9 t CO$_2$ yr^{-1} per capita (Fig. 3). That is about 20 times the

© 2018 Springer Nature Limited. All rights reserved.

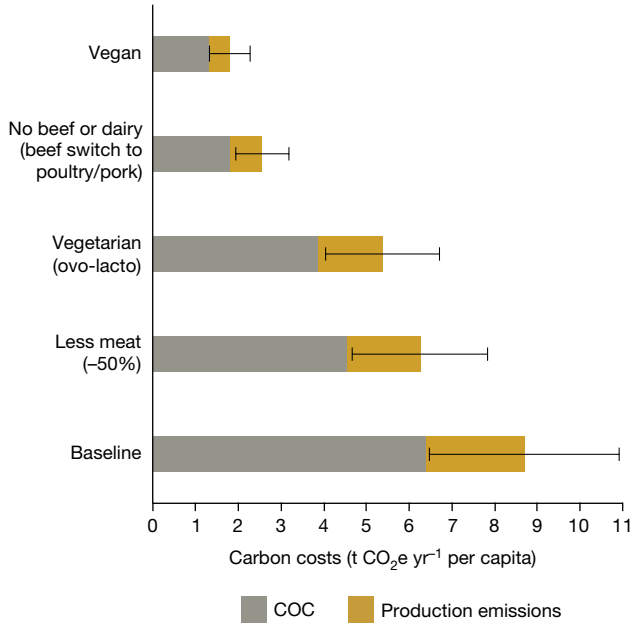


Fig. 3 | Carbon costs of different diets based on the carbon benefits index. Error bars reflect the range of literature estimates of vegetation and soil carbon stocks used in part to derive the COCs.

most-emitting diet estimate in Tilman and Clark[4] and equivalent to GHGs typically assigned to each European's consumption of all goods, including energy[23]. Shifting from beef and dairy would reduce those emissions by 70%. Although animal products offer health benefits for the food-insecure[24], we estimate much larger climate benefits than others if the wealthy consume less beef and dairy.

As these examples illustrate, our analysis finds that consumption and LUCs can have many times larger implications for climate change than often calculated. By undercounting the carbon storage opportunity costs of land, LCAs and economic models can greatly overvalue mere shifts in land uses—for example, shifting croplands or pasture to forest or bioenergy—and undervalue both increases in pasture or crop productivity and reductions in demand—such as shifts to diets low in beef and milk.

By using average global costs as a benchmark, our method evaluates the comparative carbon advantage of different land uses. Even between two inefficient land uses, the less inefficient one will generate more benefits. As yields and farming areas change, COCs must change.

Despite many estimation uncertainties (Supplementary Information), implications for policy appear to be insensitive to them. Varying COCs on the basis of native carbon estimates that are ±20% of our estimates for vegetation and ±30% for soils and changing discount rates to 2% and 6% (Supplementary Table 3) do not alter the directions of our examples (Supplementary Tables 5–9). Because of scientific uncertainties, however, our index does not incorporate biophysical effects of LUC (for example, albedo), which could be substantial.

Our index assumes that food would be replaced at global-average, rather than marginal, costs. In the real world, marginal carbon costs could differ through price effects or because a food's replacement land physically differs from the global average. (Our calculator does allow a user to select marginal COCs as a lower or higher percentage of average COCs.) We therefore suggest the following uses for our index.

First, our index can be used to evaluate shifts from agriculture to forest or bioenergy. Efforts to deliberately replace food production with forests or bioenergy for climate purposes could require sizable carbon benefits as one core criterion. Regardless of whether price effects alter consumption or the productivity of other farms, the world is unlikely to achieve the ultimate climate goals through LUCs that reduce the global capacity to store carbon. In addition, changing the production of one hectare of land to try to lower GHGs by reducing food consumption is probably inefficient and inequitable, as suggested by the SUV example. Other farmers would probably replace most of the production, and higher prices would typically depress consumption by the poor more than by the rich owing to the greater price-sensitivity of the former[25]. To reduce inefficient consumption, targeted taxes or other demand strategies could be more efficient and equitable.

Second, our index can be used in attributional LCAs to assign land-use carbon costs to consumption choices, as we show in our diet examples. Those LCAs also use average GHG costs of production, rather than some estimate of marginal costs, and similarly assume that one person's change in consumption equally alters global consumption.

Third, our index can be used as a benchmark for evaluating predictive models. Accurately projecting marginal, rather than average, consequences of one hectare's changes would still have value. Doing so requires economic models, but results greatly vary by model or assumption[3,11,26,27]. Only a small number of the demand and supply elasticities required by global models have been econometrically estimated. Missing critical estimates include almost all cross-price elasticities, almost all medium-to-long-term elasticities, and supply elasticities of different pasture systems, although pasture occupies two-thirds of all agricultural land (Supplementary Information). Our view is that global land sparing is powerful[15], although often hidden, because gains in local yields can increase competitiveness and encourage local expansion[1]. Although our index cannot by itself answer these questions, the COC provides a useful benchmark to evaluate model results. For example, California's estimates of ILUC from maize ethanol[28] are about 10% of the average global loss of carbon generated by producing the required maize (using California's amortization period and after accounting for by-products). By providing this average cost, our index helps to evaluate a model's justification for estimating greatly different marginal costs.

Last, where some conversion of natural vegetation to agriculture is inevitable, such as for oil palm in Southeast Asia[29] or for staple foods in Africa[14], our index could help to determine the most efficient lands and crops to choose. For policy reasons, however, we advise great caution in using the index to justify conversion of native vegetation based on claims of high food yields. Because climate strategies require quick elimination of emissions from LUCs, clearing land in one location does not provide a general solution, even if clearing elsewhere would be worse. Strategies to reduce LUCs require strong policies to discourage expansion, so farmers intensify instead. It may be tempting to exaggerate likely yields on lands proposed for conversion, and promises of intensive management cannot justify conversions if the same investment could generate equal yields on existing cropland. However, these kinds of conversion also have high potential to harm biodiversity and other ecosystem values, which our index does not measure and which must be evaluated separately.

Overall, the concept of 'carbon benefits' offers an alternative to the concept of 'leakage', which assumes that land benefits the climate only by sequestering carbon or producing biofuels. Our approach recognizes that all increases in efficiency generate climate benefits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0757-z.

Received: 23 March 2017; Accepted: 12 September 2018;
Published online 12 December 2018.

1. Searchinger, T. et al. *Creating a Sustainable Food Future. A Menu of Solutions to Sustainably Feed More Than 9 Billion People by 2050* (World Resources Institute, Washington, 2014).
2. Edenhofer, O. et al. in *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Edenhofer, O. et al.) 1–30 (Cambridge Univ. Press, Cambridge, 2014).
3. Valin, H. et al. The future of food demand: understanding differences in global economic models. *Agric. Econ.* **45**, 51–67 (2014).

© 2018 Springer Nature Limited. All rights reserved.

4. Tilman, D. & Clark, M. Global diets link environmental sustainability and human health. *Nature* **515**, 518–522 (2014).
5. Eshel, G., Shepon, A., Makov, T. & Milo, R. Land, irrigation water, greenhouse gas, and reactive nitrogen burdens of meat, eggs, and dairy production in the United States. *Proc. Natl Acad. Sci. USA* **111**, 11996–12001 (2014).
6. Food and Agriculture Organization of the United Nations. *EX-Ante Carbon-balance Tool (EX-ACT)* http://www.fao.org/tc/exact/ex-act-home/en/ (2017).
7. Colomb, V. et al. Selection of appropriate calculators for landscape-scale greenhouse gas assessment for agriculture and forestry. *Environ. Res. Lett.* **8**, 015029 (2013).
8. Gerber, P. et al. *Tackling Climate Change Through Livestock: A Global Assessment of Emissions and Mitigation Opportunities* (Food and Agriculture Organization of the United Nations, Rome, 2013).
9. Escobar, N., Ribal, J., Clemente, G. & Sanjuán, N. Consequential LCA of two alternative systems for biodiesel consumption in Spain, considering uncertainty. *J. Clean. Prod.* **79**, 61–73 (2014).
10. Audsley, E. et al. *How Low Can We Go? An Assessment of Greenhouse Gas Emissions from the UK Food System and the Scope Reduction by 2050* (WWF-UK, 2009).
11. Hertel, T. W. et al. Effects of US maize ethanol on global land use and greenhouse gas emissions: estimating market-mediated responses. *Bioscience* **60**, 223–231 (2010).
12. Ranganathan, J. et al. *Shifting Diets for a Sustainable Food Future* (World Resources Institute, Washington, 2016).
13. Johnson, J. A., Runge, C. F., Senauer, B., Foley, J. & Polasky, S. Global agriculture and carbon trade-offs. *Proc. Natl Acad. Sci. USA* **111**, 12342–12347 (2014).
14. Estes, L. D. et al. Reconciling agriculture, carbon and biodiversity in a savannah transformation frontier. *Phil. Trans. R. Soc. Lond. B* **371**, 20150316 (2016).
15. Searchinger, T. D., Edwards, R., Mulligan, D., Heimlich, R. & Plevin, R. Do biofuel policies seek to cut emissions by cutting food? *Science* **347**, 1420–1422 (2015).
16. Strassburg, B. B. N. et al. When enough should be enough: improving the use of current agricultural lands could meet production demands and spare natural habitats in Brazil. *Glob. Environ. Change* **28**, 84–97 (2014).
17. Cardoso, A. S. et al. Impact of the intensification of beef production in Brazil on greenhouse gas emissions and land use. *Agric. Syst.* **143**, 86–96 (2016).
18. Luyssaert, S. et al. CO_2 balance of boreal, temperate, and tropical forests derived from a global database. *Glob. Change Biol.* **13**, 2509–2537 (2007).
19. Xu, B., Yang, Y., Li, P., Shen, H. & Fang, J. Global patterns of ecosystem carbon flux in forests: a biometric data-based synthesis. *Glob. Biogeochem. Cycles* **28**, 962–973 (2014).
20. Hudiburg, T. W., Davis, S. C., Parton, W. & Delucia, E. H. Bioenergy crop greenhouse gas mitigation potential under a range of management practices. *Glob. Change Biol. Bioenergy* **7**, 366–374 (2015).
21. Harris, Z. M., Spake, R. & Taylor, G. Land use change to bioenergy: a meta-analysis of soil carbon and GHG emissions. *Biomass Bioenergy* **82**, 27–39 (2015).
22. Bryngelsson, D., Wirsenius, S., Hedenus, F. & Sonesson, U. How can the EU climate targets be met? A combined analysis of technological and demand-side changes in food and agriculture. *Food Policy* **59**, 152–164 (2016).
23. Moll, S. & Remond-Tiedrez, I. *CO_2Emissions Induced by EU's Final Use of Products are Estimated to Be 9 Tonnes per Capita* (Eurostat, Luxembourg, 2011).
24. Neumann, C. G., Demment, M. W., Maretzki, A., Drorbaugh, N. & Galvin, K. A. in *Livestock in a Changing Landscape, Volume 1: Drivers, Consequences, and Responses* (eds Steinfeld, H. et al.) 221–248 (Island Press, Washington, 2010).
25. Regmi, A., Deepak, M. S., Seale, J. L. Jr & Bernstein, J. in *Changing Structure of Global Food Consumption and Trade* (ed. Regmi, A.) 14–22 (United States Department of Agriculture Economic Research Service, Washington, 2001).
26. Prestele, R. et al. Hotspots of uncertainty in land use and land cover change projections: a global scale model comparison. *Glob. Change Biol.* **22**, 3967–3983 (2016).
27. Plevin, R., O'Hare, M., Jones, A., Torn, M. & Gibbs, H. Greenhouse gas emissions from biofuels' indirect land use change are uncertain but may be much greater than previously estimated. *Environ. Sci. Technol.* **44**, 8015–8021 (2010).
28. California Air Resources Board. *Low Carbon Fuel Standard. Final Regulation Order* Table 5, https://www.arb.ca.gov/regact/2015/lcfs2015/lcfsfinalregorder.pdf (2015).
29. Corley, R. H. V. How much palm oil do we need? *Environ. Sci. Policy* **12**, 134–139 (2009).

Acknowledgements We thank the David and Lucille Packard Foundation and the Norwegian Agency for Development Cooperation for financial support. We thank L. Germer for work on programming the Carbon Benefits Calculator, and J. Moretti and C. Klirs for help with graphics. We thank R. Boddey and A. Cardoso for additional data, advice and citations necessary for the Brazil examples. We thank R. Conant, K. Erb and H. Haberl for contributions to early thinking for this project, and C. Malins, S. Yeh and M. O'Hare for comments on early drafts.

Reviewer information *Nature* thanks L. Firbank and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed to all features of the paper. In addition, T.D.S. conceived the idea for the research and wrote the paper. S.W. programmed our spreadsheet model. T.B. performed global vegetation modelling and spatial analysis. P.D. estimated product-specific PEMs.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0757-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0757-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.D.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Calculating COCs using the carbon loss method. *Basic method for crops.* The COC for each crop equals the aggregate, time-discounted carbon lost from native vegetation and soils on land used to produce the crop globally (the numerator) in kilograms, divided by the aggregate, time-discounted annual production for that crop in kilograms. The result is multiplied by 3.67 to be expressed as kilograms of CO_2e per kilogram of crop.

To estimate native carbon stocks of both vegetation and soils of existing cropland (Extended Data Figs. 3, 4), we employ a combination of vegetation modelling and biome estimates. We use the Lund–Potsdam–Jena managed land (LPJmL) dynamic global vegetation model (DGVM)[30,31] to estimate native carbon stock in each 0.5° × 0.5° grid cell, but we scale the LPJmL results in each pixel so the average biome values of our adjusted LPJmL results match those of average reference values for the biome from the literature[32–36]. After analysing estimates of native carbon stocks available from three other DGVMs and determining that many of their estimates were too implausible to use (as discussed extensively in Supplementary Information), we chose LPJmL because its average estimates at the biome level match literature estimates fairly well.

Although some previous efforts used average carbon stocks for entire biomes to estimate stocks for each crop[13,37], these biomes are large and include lands with very different native carbon stocks and productivities used differently by different crops; so they cannot properly distinguish among crops. Our method uses empirical measures to adjust our DGVM results at the biome level to match empirical estimates but preserves the higher spatial resolution of the DGVM, so that carbon stocks within a biome can reflect major physical differences, such as rainfall. This method also implicitly incorporates the effects of disturbance, for example, from fire and wind, because they are automatically incorporated into biome estimates.

To identify the locations of different crops, we use maps provided by the Spatial Production Allocation Model (SPAM) for 42 crops in the year 2005[38] and estimate carbon losses on all cropland used by each crop separately. The loss represents the difference between vegetative carbon in native vegetation (including both above- and below-ground parts) and the average carbon stock of the crop. For crop carbon stocks, we use data from ref. [39] for perennial crops. For annual crops we assume annual average carbon stocks of 25% of the peak values and a whole-plant multiplier of 2.5 from the carbon in harvested crops. For conversion to cropland, we also assume loss of 25% of soil carbon within the top metre of soils, consistent both with several other global studies[39,40] and with a range of new meta-analyses[41–45]. For global crop production, we use data from the FAOSTAT database[46]. For products that are only a portion of crops, such as vegetable oils, we apportion crop output based on energy content.

Time discounting. Conversion of land from forest to cropland loses carbon relatively quickly whereas the benefits of crop production for food or bioenergy extend over time. To reflect the values of earlier emissions reductions, we apply a discount rate both to the stream of carbon losses in the numerator of the COC and to the stream of production in the denominator. For vegetation losses in the numerator, we estimate the annual stream of losses using exponential decay functions from ref. [47], which vary by type of vegetation and climatic regions. For soil carbon losses we consider these rates[47] to be too fast, and instead follow the exponential carbon response function from ref. [48]. For the conversion of forests to cropland, it implies the loss of 98% of the volatile soil organic carbon (SOC) stock (25% of the SOC in the upper 1 m of the soil) within 20 years and is therefore consistent with the default period of the Intergovernmental Panel on Climate Change (IPCC). We apply similar discounting to the stream of crop production in the denominator.

In our base case, we use a 4% discount rate over 100 years for reasons that we explore more thoroughly in Supplementary Information. The choice of discount rate should be solely a question of climate policy for valuing mitigation over time, reflecting, among other matters, the cost of short-term as well as long-term damages, risks of crossing thresholds, and the time value of money. In general, a 4% discount rate is consistent with a 4% real return on investment[49] and a constant cost of a tonne of emissions over time. It also produces results roughly equivalent to the implicit treatment of time discounting by USA federal and California biofuel policies, which use a 30-year amortization period for carbon lost from land conversion owing to biofuels.

Calculating carbon loss from organic soils. Because the LPJmL model does not include detailed representations of peatland development and distribution, we use a global map of peatland regions[50] to estimate emissions from organic soils under croplands. We determine shares of peatland soils for all SPAM crop distribution maps[38] and apply emission factors from ref. [51] (using a rate of 15 t C ha^{-1} yr^{-1} for oil palm). We also assume 8 Mha of drained peatland for pasture[52] and emission rates equal to half of cropland emission rates because of lower need for drainage.

Calculating COCs using the carbon gain method. To calculate COCs with the carbon gain method, we assume that if increasing yields result in a reduction in agricultural land, the productive potential of the land no longer used for crops can be restored to forest. We base this productive potential on the net primary productivity of the native vegetation (NPP_{nat}) of global hectares devoted to each crop, expressed in tonnes of carbon per hectare per year (t C ha^{-1} yr^{-1}). Although land management can increase or decrease this productive potential, this native productivity (which reflects rainfall, solar radiation, temperature and soil type, among other factors) provides a reasonable measure of inherent productive potential. We use LPJmL to estimate and map NPP_{nat} (Extended Data Fig. 5) and then use the SPAM 2005 v3.2 cropland maps to estimate the average NPP_{nat} of the land used for each type of crop. This average NPP_{nat} per hectare, divided by the average yield of that crop, generates the amount of NPP_{nat} used to produce a tonne of each crop, which can be converted to kilograms CO_2 per kilogram of crop.

To determine how much potential carbon sequestration would be generated on average by devoting one tonne of NPP_{nat} to reforestation, we need to determine both the NPP_{nat} of forests that have become cropland and the average carbon sequestration rate on croplands modified to regenerate forest. We estimate the average NPP_{nat} of all tropical croplands that were originally forest to be 9.7 t C ha^{-1} yr^{-1}. We then use the mean of three recent meta-analyses of carbon fluxes in forests to estimate average carbon sequestration in regenerating tropical forests[18,19,53,54] over 100 years at 4.1 t C ha^{-1} yr^{-1} in vegetation and soils. Dividing the output (4.1 t C ha^{-1} yr^{-1}) by NPP_{nat} produces a ratio of 0.42 t C ha^{-1} yr^{-1} for every tonne of NPP_{nat} available, which equals 1.5 kg CO_2 (kg CO_2 NPP_{nat}) $^{-1}$ for the tropics.

Extending our analysis to originally forested croplands in both the tropics and the temperate zone, we estimate the NPP_{nat} at 8.5 t C ha^{-1} yr^{-1} from LPJmL, and the average annual carbon sequestration rate for regrowing forests[18,19,54] is 3.6 t C ha^{-1} yr^{-1}. Although both figures are lower than those obtained for the tropics alone, they generate the same ratio of 0.42 t C ha^{-1} yr^{-1} sequestration for every tonne of NPP_{nat} available. As the vast majority of the world's croplands are located in temperate to tropical regions[55], we use this benchmark.

For each crop, the COC calculated using this method equals that crop's ratio of NPP_{nat} to crop output in kilogram of CO_2 per kilogram of crop, multiplied by this 1.5 kg CO_2 (kg CO_2 NPP_{nat})$^{-1}$, which generates kilograms of CO_2 per kilogram of crop.

Calculating COCs of livestock products. The global-average COC of livestock products (meat, dairy and eggs) equals the global-average COC of feeds, including portions of crops, such as oilseed meals, used to produce them. We estimate the global-average feed use per unit of livestock output based on the few publications available on global feed use[56–58]. We calibrated these data against FAOSTAT data on forage production and FAOSTAT feed use data for cereals, tubers, oil crops, pulses, brans, molasses and oil meals, so our total global feed use equals that in the FAOSTAT data. We treated fibrous, low-value by-products, such as crop residues and straw, as land-free sources of feed (that is, they have no COC), which applies to roughly 20% of global feed use in dry matter[56].

Because ruminants heavily rely on grasses, we estimate the COC of permanent grazing land and apportion this COC to the forage from permanent grasslands for beef, bovine milk, and mutton based on global estimates of their relative consumption of grasses[58,59].

To estimate carbon losses on pasture, we use the HYDE 3.2 land-use map[60], which estimates 2.8 billion hectares of grazing land. We overlay pastures with our estimate of native vegetation carbon stocks described above. Changes in SOC following the conversion of forests or grasslands into pastures remain disputed. Because effects in the tropics vary from negative to positive depending on grazing practices[42,61,62], we assume no change in soil carbon for tropical pastures. Relying on a recent meta-analysis for temperature pastures, we assume a 10% loss of carbon in temperate pastures[63].

For grazing lands that were naturally grassland (tree canopy cover less than 10%), we also assume no loss of vegetative carbon. For grazing lands that were naturally forested (more than 60% tree cover), we estimate a loss of all tree carbon and replacement by grass carbon assuming that such areas must be cleared to show up as grasslands in land cover classifications from satellite data. For grazing lands that were naturally some kind of woody savanna, we assume 75% loss of vegetation carbon for woody savannas (30%–60% canopy cover) and 50% loss of vegetation carbon for savannas (10%–30% canopy cover), also based on assumptions about satellite data. On average for all grassland, these assumptions imply a 92% loss of vegetation carbon. Because this carbon loss is dominated by the loss of dense forests, assumptions for carbon losses on native grasslands and woody savannas have little consequence (see Supplementary Table 2).

Calculating COCs using the carbon gain method. For the carbon gain method, we estimate the NPP_{nat} of grazing lands using LPJmL. Because grazing lands maintain native vegetation carbon stocks to a varying degree, we assign some NPP_{nat} to the maintenance of these carbon stocks using complementary numbers to those for vegetation carbon loss (see previous paragraph). Hence, for natural grasslands we assume 100% for forests, 25% for woody savannas, 50% for less-woody savannas and 0% for grasslands that were originally forests.

We follow the same approach to time-discounting COCs for livestock and pasture feeds as described above for crops.

© 2018 Springer Nature Limited. All rights reserved.

Estimating PEMs for crops and crop products. We estimate the global-average PEMs for each crop for which we derived COCs based on sources (listed below) that ultimately rely on the IPCC Tier 1 or Tier 2 methods. These emissions were built into a global agriculture emissions accounting model (GlobAgri-WRR; developed by CIRAD, the World Resources Institute, Princeton University and INRA (Institut National de la Recherche Agronomique) for the World Resources Report of the World Resources Institute[12]) that uses the same methodology for agricultural product balances and similar livestock data as those used in studies by CIRAD and INRA[64,65]. Although this model contains many other features, for the purposes of this study, the model essentially provides a spreadsheet that adds up the emissions in the production process for crops in each region where they are produced and divides them by the total production. Sources of emissions are as follows.

Emissions from nitrogen use. Nitrogen balance, harvested nitrogen, nitrogen fixation and use of fixed nitrogen, in addition to legumes needs of following crops, are based on data used in the analysis of ref. [59], with manure nitrogen rescaled using data from ref. [58]. Emissions from nitrogen in the form of nitrous oxide are based on IPCC Tier 1 emission factors for direct and indirect emissions. Emissions from the manufacture and transport of nitrogen are based on analysis by the US Environmental Protection Agency (EPA)[66]. To compute N_2O nitrogen residue emissions, we apply a factor of N_2O emissions per harvested nitrogen, obtained by dividing the FAOSTAT total residue N_2O emission by the total harvested nitrogen for each country.

Rice methane. Rice methane emissions rates are based on a spreadsheet model[67], adjusted to match expert opinions of mid-season drainage or multiple drainages[68].

Emissions from potash and phosphorus consumption. Quantities of potash and phosphorus used per crop are based on estimates for 2003 and 2007 data originally compiled by the International Fertilizer Institute and completed by FertiStat[69]. We use methods described in Supplementary Information to estimate application rates for crops not represented in the initial data. Country-level fertilizer consumption from FAOSTAT is then used to rescale over time the rates per crop per unit area. Emissions are based on estimates of those associated with phosphate and potash extraction in the analysis of EPA[66].

Pesticides emissions. Pesticide quantities are taken from FAOSTAT and emissions per kilogram of active ingredients in pesticides are based on the analysis of EPA[66].

Direct on-farm energy use. Emissions for energy used directly on farms are taken from FAOSTAT[46]. To allocate emissions to individual crops, we first deduct a global number for livestock PEMs based on previous estimates[8] and on the hypothesis of a constant coefficient for emissions per energy content of livestock product. Then we allocate the remainder to crops using professional judgment supported by different lifecycle calculations.

We use 100-year global warming potentials of 298 for N_2O and 34 for CH_4 based on recommendations in the latest assessment report by the IPCC[70].

Estimating PEMs for livestock products. Although GlobAgri-WRR estimates livestock PEMs that rely heavily on ref. [58], we do not use GlobAgri-WRR for this purpose, in part because it uses the ruminant model to estimate methane emissions from enteric fermentation, which is not easily accessible by others for use in estimating methane emissions for individual farms. We therefore use IPCC Tier 2 methods to estimate methane from enteric fermentation based on the feed use estimates obtained in this study. Non-enteric livestock emission sources are estimated on the basis of GLEAM model results[8,71,72]. To be consistent with our crop production estimates, PEMs for feed (which contribute to livestock PEMs) are based on GlobAgri-WRR, as described above. For emissions of nitrous oxide from pasture, we use estimates of nitrogen applied to pasture generated from ref. [59].

Biofuel and biofuel by-product COCs and PEMs for estimating GHG costs of consumption. As in the case of livestock products, the global-average COC of biofuels equals the global-average COC of the feedstock (that is, crop products) used to produce them. Process yields and GHG emission data are based on ref. [73], except in the case of grass-based ethanol, where they are based on ref. [74], which assumed conversion of biomass to ethanol of 375 litres per tonne of dry matter, which we use as well. In both studies[73,74], the GHG savings due to electricity by-products from sugarcane or cellulosic ethanol production are allocated to the biofuel, which reduces its nominal PEMs. To account for the land and GHG-sparing value of feed co-products from maize and wheat ethanol (distillers' dried grains with solubles, DDGS) we use the substitution method. We estimate the specific crops that the DDGS would replace based on ref. [75]. We then apply the COC values of the crops substituted. (Despite uncertainty about DDGS uses, the analysis[75] generally values DDGS for its protein value, which increases by-product values compared to use for calories.)

Equation for calculating carbon benefits of production. The carbon benefits (CB; in kg CO_2e ha^{-1} yr^{-1}) are calculated as

$$CB = COC_s + PEM_{bfits} + CARBST_{ch} + FOS_{sav} \quad (1)$$

with

$$COC_s = \boldsymbol{Y} \cdot \mathbf{COC}$$
$$PEM_{bfits} = \boldsymbol{Y} \cdot (\mathbf{PEM}_{avg} - \mathbf{PEM}_h)$$
$$CARBST_{ch} = \frac{PDV_{cs\text{-}ch}}{PDV_{ha\text{-}yr_{1\to 100}}}$$
$$FOS_{sav} = \mathbf{BIOFY} \cdot (\mathbf{FOSEF} - \mathbf{BIOFEF})$$

COC_s is the total COC (kg CO_2e ha^{-1} yr^{-1}), $\boldsymbol{Y}$ is a vector of yield(s) of agricultural product(s) (including any biofuel feed by-products; kg product ha^{-1} yr^{-1}) and **COC** is a vector of the COC(s) of agricultural product(s) (kg CO_2e per kg product). PEM_{bfits} is the total benefits (or costs) from PEMs on the hectare (kg CO_2e ha^{-1} yr^{-1}), where $\mathbf{PEM}_{avg}$ is a vector of the global-average PEMs for each agricultural product (kg CO_2e per kg product) and $\mathbf{PEM}_h$ is the PEMs on the hectare analysed (kg CO_2e per kg product). This equation applies to all crop and animal outputs other than biofuels and includes biofuel by-products used for feed or food, whose value is based on the crop products that they substitute. $CARBST_{ch}$ is the time-discounted benefit from the annual change in carbon storage in vegetation and soils (kg CO_2e ha^{-1} yr^{-1}) where $PVD_{cs\text{-}ch}$ is the present discounted value of the expected change in carbon storage (kg CO_2e) and $PDV_{ha\text{-}yr_{1\to 100}}$ is the present discounted value of each hectare, each year, over 100 years (ha yr). (Discounting the stream of hectare years is equivalent for carbon-stock changes to discounting the stream of crop production in the denominator of the COC.) FOS_{sav} is the total fossil fuel savings (net; kg CO_2e ha^{-1} yr^{-1}), where **BIOFY** is a vector of biofuel yield(s) (MJ ha^{-1} yr^{-1}), **FOSEF** is a vector of fossil fuel emission factor(s) (production and combustion, but not LUC; kg CO_2e MJ^{-1}) and **BIOFEF** is a vector of biofuel PEMs factor(s) (production only, not combustion or LUC; kg CO_2e MJ^{-1}). **BIOFEF** is partially a function of the agricultural practices on a particular parcel of land and partially a function of the emissions involved downstream in the conversion and transportation processes. When evaluating biofuels, the PEM applied to the biofuel by-product should allocate total farm PEMs between that by-product and the biofuel.

Equation for calculating carbon costs of consumption. The carbon cost of consumption (CCC; kg CO_2e) is

$$CCC = \mathbf{CONSUM} \cdot (\mathbf{COC} + \mathbf{PEM}) \quad (2)$$

where **CONSUM** is the consumption of a product(s) in kilograms, and the COC and PEM of each product is expressed in kilograms CO_2e per kilogram product.

Sensitivity calculations. We perform sensitivity analysis for COCs by varying soil and vegetation carbon estimates across all areas. We used an uncertainty of ±30% for soil carbon, based on a review of differential soil carbon estimates[76], and ±20% for vegetative carbon, based on our assumption that their uncertainties are substantially lower. We generate high and low COCs based on this range and on alternative discount rates of 2% and 6% (Supplementary Table 3). We show the effects of these assumptions on all examples analysed here in Supplementary Tables 5–9. The results show no directional changes in land-use comparisons, except for a few uses that have very similar carbon benefits or carbon costs in our central scenario and that can shift modestly from one side to another.

Uncertainties and certain data advantages of our approach. Our approach has one inherent technical advantage over modelling approaches that attempt to estimate likely carbon losses from land conversion by estimating the precise locations where land conversion will occur. To estimate where conversion will occur and the resulting carbon losses, such approaches require overlapping multiple spatial datasets, each of which has its own random errors. Even maps of cropland versus other lands have large discrepancies and substantial errors[77,78]. Overlapping the maps will produce errors wherever any individual map has errors; for example, a correct yield estimate combined with an incorrect carbon estimate will generate an incorrect result. More problematically, many cells with such errors will probably stand out as most likely or beneficial for conversion because of such errors.

Because the carbon benefits index estimates carbon loss per kilogram of crop by averaging critical parameters across all cells devoted to each crop worldwide—although opportunities for systematic errors remain—the method provides many opportunities to average out random errors. At the same time, although the COC is based on this global average, the user can use better site-specific information about the precise parcel undergoing change. Supplementary Information contains a fuller discussion of additional uncertainties.

Code availability. The carbon benefits index model, which shows the calculation of COCs and PEMs, is available for download from Pangea at https://doi.org/10.1594/PANGAEA.877266. The LPJmL model code is available at https://github.com/PIK-LPJmL/LPJmL. The Carbon Benefits Calculator, which facilitates calculation of carbon benefits using COCs and PEMs for specific parcels of land,

© 2018 Springer Nature Limited. All rights reserved.

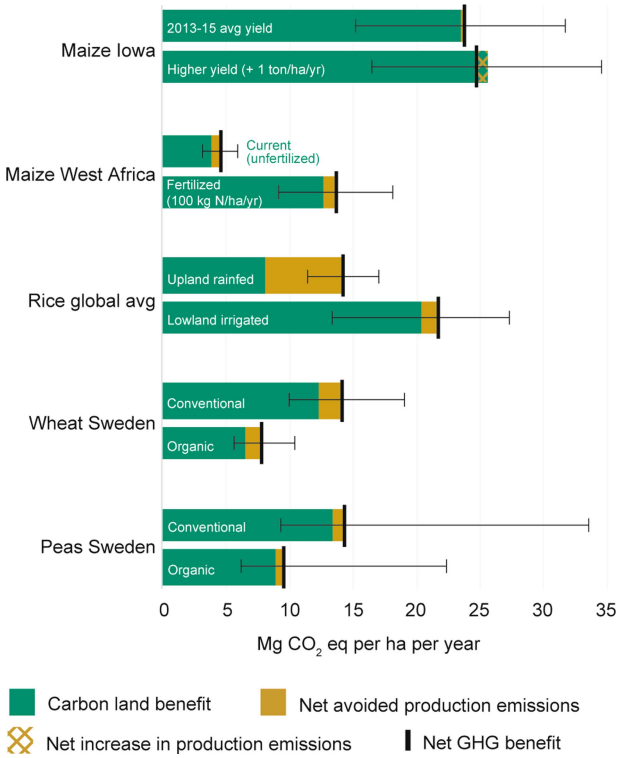
is included as Supplementary Data, and future revisions will be available at https://www.princeton.edu/~tsearchi.

Data availability

LPJmL modelling results, in the form of global carbon and native net primary productivity maps, are available at https://doi.org/10.1594/PANGAEA.877266. The different datasets used to run LPJmL for this study are publicly available and described in Supplementary Information along with links. Any other materials generated for this study are available from the corresponding author on reasonable request.

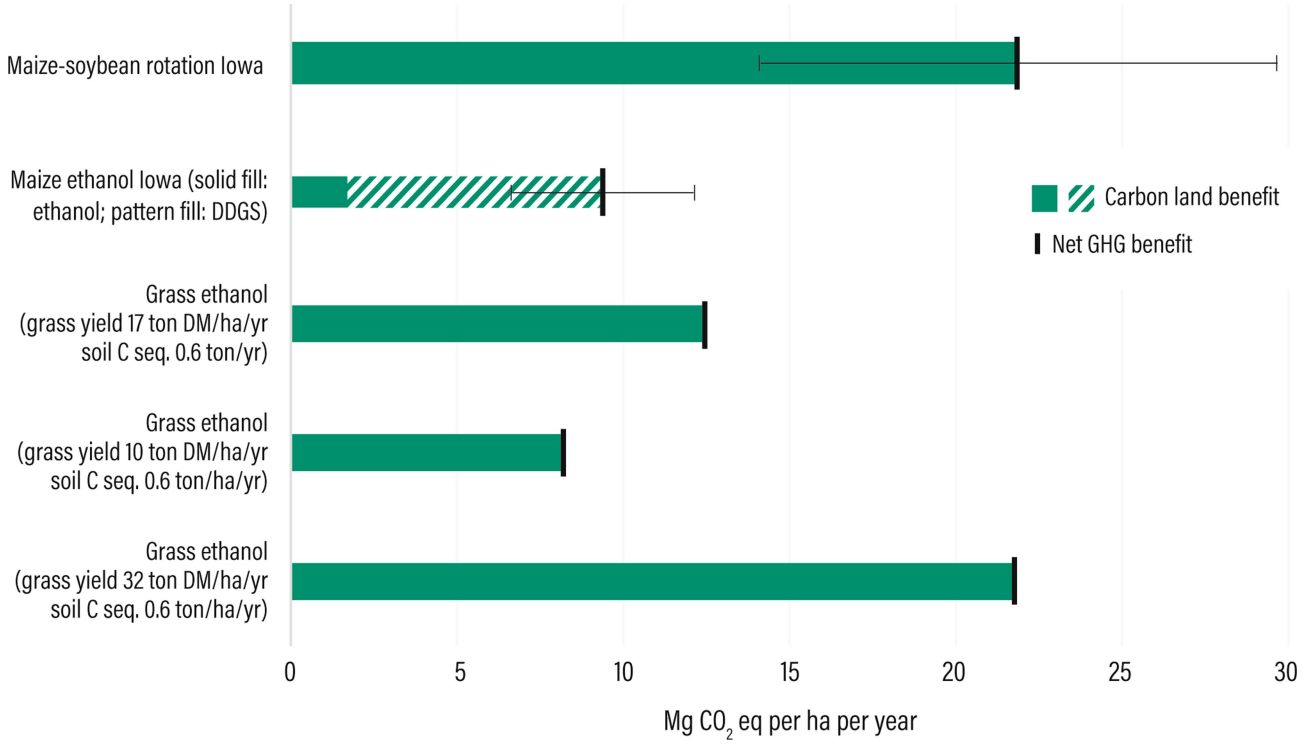
30. Sitch, S. et al. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Glob. Change Biol.* **9**, 161–185 (2003).
31. Bondeau, A. et al. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Glob. Change Biol.* **13**, 679–706 (2007).
32. Eggleston, H. S., Buendia, L., Miwa, K., Ngara, T. & Tanabe, K. (eds) *2006 IPCC Guidelines for National Greenhouse Gas Inventories* (IGES, Hayama, 2006).
33. Jobbágy, E. G. & Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **10**, 423–436 (2000).
34. Malhi, Y., Baldocchi, D. D. & Jarvis, P. G. The carbon balance of tropical, temperate and boreal forests. *Plant Cell Environ.* **22**, 715–740 (1999).
35. Roy, J., Saugier, B. & Mooney, H. A. (eds) *Terrestrial Global Productivity* (Academic Press, San Diego, 2001).
36. Trumper, K. et al. *The Natural Fix? The role of Ecosystems in Climate Mitigation* (UNEP-WCMC, Cambridge, 2009).
37. West, P. C. et al. Trading carbon for food: global comparison of carbon stocks vs. crop yields on agricultural land. *Proc. Natl Acad. Sci. USA* **107**, 19645–19648 (2010).
38. You, L. et al. *Spatial Production Allocation Model (SPAM) 2005 v3.2* http://MapSPAM.info (2017).
39. Gibbs, H. K. et al. Carbon payback times for crop-based biofuel expansion in the tropics: the effects of changing yield and technology. *Environ. Res. Lett.* **3**, 034001 (2008).
40. Searchinger, T. D. et al. Use of U.S. croplands for biofuels increases greenhouse gases through emissions from land-use change. *Science* **319**, 1238–1240 (2008).
41. Guo, L. B. & Gifford, R. M. Soil carbon stocks and land use change: a meta analysis. *Glob. Change Biol.* **8**, 345–360 (2002).
42. Don, A., Schumacher, J. & Freibauer, A. Impact of tropical land-use change on soil organic carbon stocks – a meta-analysis. *Glob. Change Biol.* **17**, 1658–1670 (2011).
43. Wei, X., Shao, M., Gale, W. & Li, L. Global pattern of soil carbon losses due to the conversion of forests to agricultural land. *Sci. Rep.* **4**, 4062 (2014).
44. Sanderman, J., Hengl, T. & Fiske, G. J. Soil carbon debt of 12,000 years of human land use. *Proc. Natl Acad. Sci. USA* **114**, 9575–9580 (2017); correction **115**, E1700 (2018).
45. Nyawira, S. S., Nabel, J. E. M. S., Don, A., Brovkin, V. & Pongratz, J. Soil carbon response to land-use change: evaluation of a global vegetation model using observational meta-analyses. *Biogeosciences* **13**, 5661–5675 (2016).
46. Food and Agriculture Organization of the United Nations. *FAOSTAT Database* http://www.fao.org/faostat/en/#home (2016).
47. Anderson-Teixeira, K. J. & DeLucia, E. H. The greenhouse gas value of ecosystems. *Glob. Change Biol.* **17**, 425–438 (2011).
48. Poeplau, C. et al. Temporal dynamics of soil organic carbon after land-use change in the temperate zone – carbon response functions as a model approach. *Glob. Change Biol.* **17**, 2415–2427 (2011).
49. Nordhaus, W. D. Revisiting the social cost of carbon. *Proc. Natl Acad. Sci. USA* **114**, 1518–1523 (2017).
50. Yu, Z., Loisel, J., Brosseau, D. P., Beilman, D. W. & Hunt, S. J. Global peatland dynamics since the Last Glacial Maximum. *Geophys. Res. Lett.* **37**, L13402 (2010).
51. Hiraishi, T. et al. (eds) *2013 Supplement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories: Wetlands* (IPCC, Geneva, 2014).
52. Biancalani, R. & Avagyan, A. *Towards Climate-Responsible Peatlands Management* (Food and Agriculture Organization of the United Nations, Rome, 2014).
53. Poorter, L. et al. Biomass resilience of Neotropical secondary forests. *Nature* **530**, 211–214 (2016).
54. Anderson-Teixeira, K. J., Wang, M. M. H., McGarvey, J. C. & LeBauer, D. S. Carbon dynamics of mature and regrowth tropical forests derived from a pantropical database (TropForC-db). *Glob. Change Biol.* **22**, 1690–1709 (2016).
55. Ramankutty, N., Evan, A. T., Monfreda, C. & Foley, J. A. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Glob. Biogeochem. Cycles* **22**, GB1003 (2008).
56. Wirsenius, S. *Human Use of Land and Organic Materials. Modeling the Turnover of Biomass in the Global Food System*. PhD Thesis, Chalmers Univ. of Technology and Göteborg Univ. (2000).
57. Bouwman, A. F., Van der Hoek, K. W., Eickhout, B. & Soenario, I. Exploring changes in world ruminant production systems. *Agric. Syst.* **84**, 121–153 (2005).
58. Herrero, M. et al. Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems. *Proc. Natl Acad. Sci. USA* **110**, 20888–20893 (2013).
59. Zhang, X. et al. Managing nitrogen for sustainable development. *Nature* **528**, 51–59 (2015).
60. Klein Goldewijk, K.K., Beusen, A., Doelman, J. & Stehfest, E. Anthropogenic land use estimates for the Holocene – HYDE 3.2. *Earth Syst. Sci. Data* **9**, 927–953 (2017).
61. Stahl, C. et al. Soil carbon stocks after conversion of Amazonian tropical forest to grazed pasture: importance of deep soil layers. *Reg. Environ. Change* **16**, 2059–2069 (2016).
62. Fujisaki, K. et al. From forest to cropland and pasture systems: a critical review of soil organic carbon stocks changes in Amazonia. *Glob. Change Biol.* **21**, 2773–2786 (2015).
63. Zhou, G. et al. Grazing intensity significantly affects belowground carbon and nitrogen cycling in grassland ecosystems: a meta-analysis. *Glob. Change Biol.* **23**, 1167–1179 (2017).
64. Le Mouël, C. *Agrimonde-Terra Foresight: Land Use and Food Security in 2050* (CIRAD/INRA, Paris, 2016).
65. Le Mouël, C. et al. *Le Système Agricole et Alimentaire de la Région Afrique du Nord–Moyen-Orient à L'Horizon 2050 : Projections de Tendance et Analyse de Sensibilité* (INRA, Paris/Rennes, 2015).
66. EPA. *Life-Cycle Analysis of Greenhouse Gas Emissions from Renewable Fuels* (2010).
67. Yan, X., Akiyama, H., Yagi, K. & Akimoto, H. Global estimations of the inventory and mitigation potential of methane emissions from rice cultivation conducted using the 2006 Intergovernmental Panel on Climate Change Guidelines. *Glob. Biogeochem. Cycles* **23**, GB2002 (2009).
68. Adhya, T. K., Linquist, B., Searchinger, T. D., Wassmann, R. & Yan, X. *Wetting and Drying: Reducing Greenhouse Gas Emissions and Saving Water from Rice Production* (World Resources Institute, Washington DC, 2014).
69. Food and Agriculture Organization of the United Nations. *FertiStat* http://www.fao.org/tempref/agl/agll/docs/fertusebycrop.xls (2006; accessed 2016).
70. Myhre, G. et al. in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. et al.) 659–740 (Cambridge Univ. Press, Cambridge, 2013).
71. Opio, C. et al. *Greenhouse Gas Emission from Ruminant Supply Chains: A Global Life Cycle Assessment* (Food and Agriculture Organization of the United Nations, Rome, 2013).
72. MacLeod, M. et al. *Greenhouse Gas Emissions from Pig and Chicken Supply Chains: A Global Life Cycle Assessment* (Food and Agriculture Organization of the United Nations, Rome, 2013).
73. Edwards, R., Larivé, J. F., Rickeard, D. & Weindorf, W. *Well-to-Tank Report Version 4.a: JEC Well-to-Wheels Analysis* (European Commission, Joint Research Centre, Ispra, 2014).
74. Evans, S. G., Ramage, B. S., DiRocco, T. L. & Potts, M. D. Greenhouse gas mitigation on marginal land: a quantitative review of the relative benefits of forest recovery versus biofuel production. *Environ. Sci. Technol.* **49**, 2503–2511 (2015).
75. Klasing, K. C. *Displacement ratios for US corn DDGS* (International Council on Clean Transportation, Washington, 2012).
76. Scharlemann, J. P., Tanner, E. V., Hiederer, R. & Kapos, V. Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Manag.* **5**, 81–91 (2014).
77. Estes, L. et al. A large-area, spatially continuous assessment of land cover map error and its impact on downstream analyses. *Glob. Change Biol.* **24**, 322–337 (2017).
78. Fritz, S. et al. Highlighting continued uncertainty in global land cover maps for the user community. *Environ. Res. Lett.* **6**, 044005 (2011).

© 2018 Springer Nature Limited. All rights reserved.



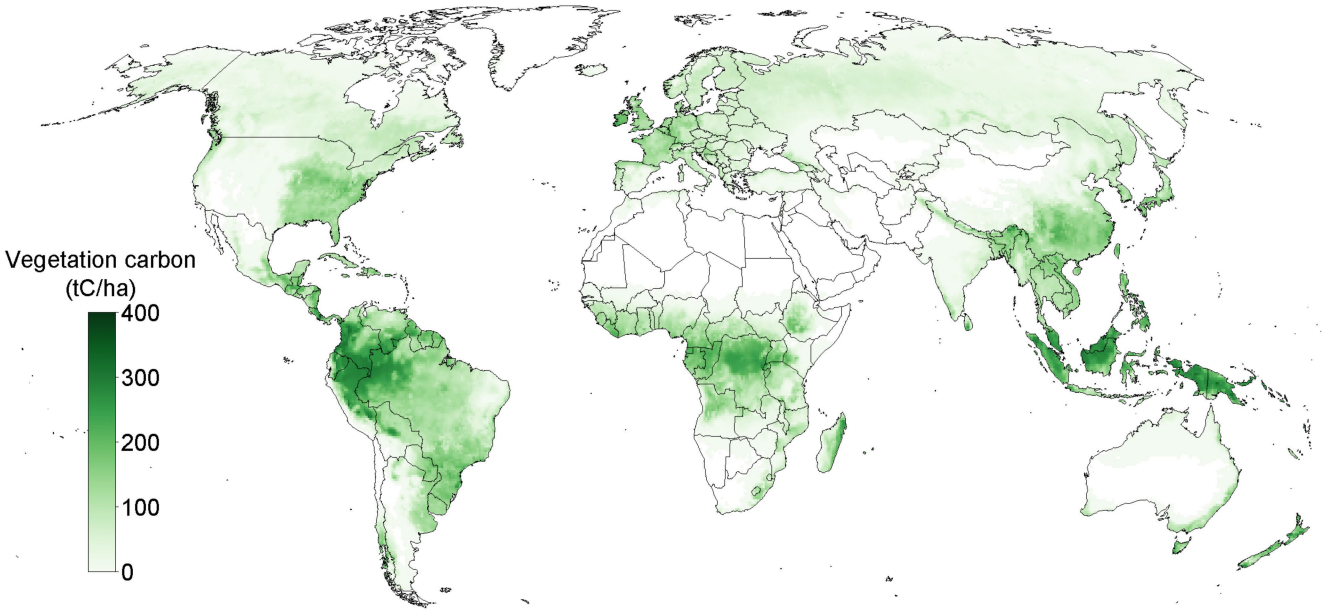
Extended Data Fig. 1 | Carbon benefits of different crop production systems based on the carbon benefits index. Error bars reflect the range of literature estimates of vegetation and soil carbon stocks.

© 2018 Springer Nature Limited. All rights reserved.



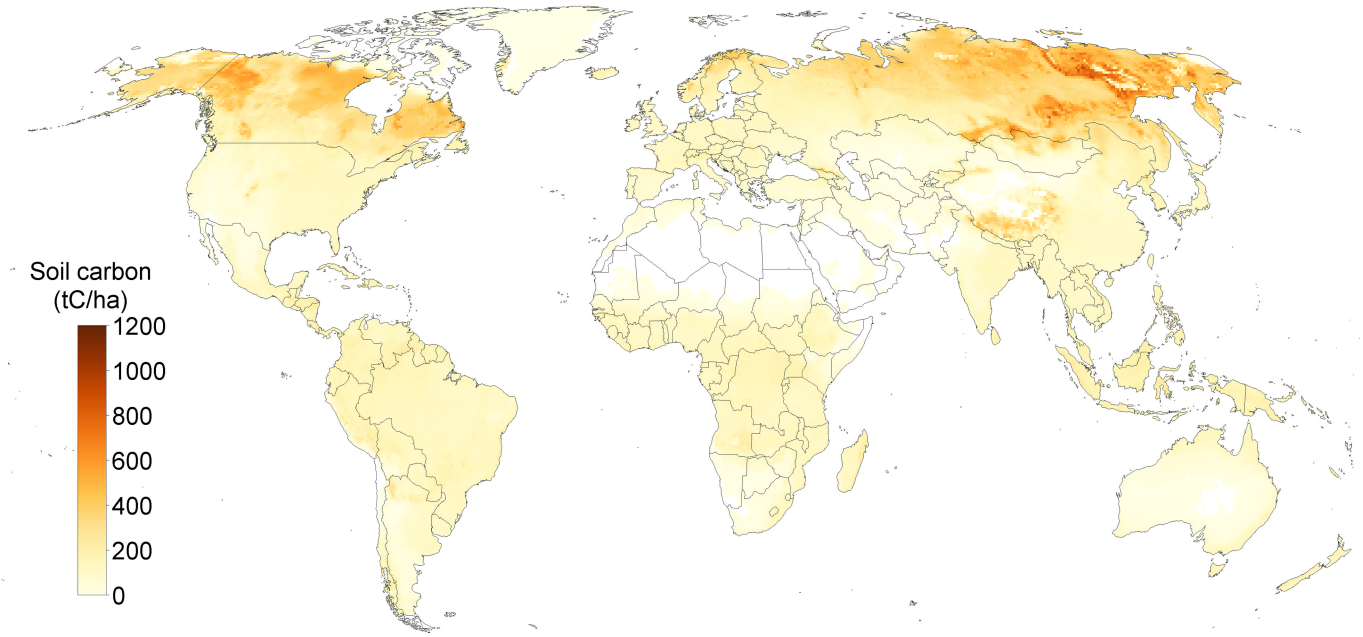
Extended Data Fig. 2 | Carbon benefits of different potential Iowa cropland uses based on the carbon benefits index. Error bars reflect the range of literature estimates of vegetation and soil carbon stocks.

© 2018 Springer Nature Limited. All rights reserved.



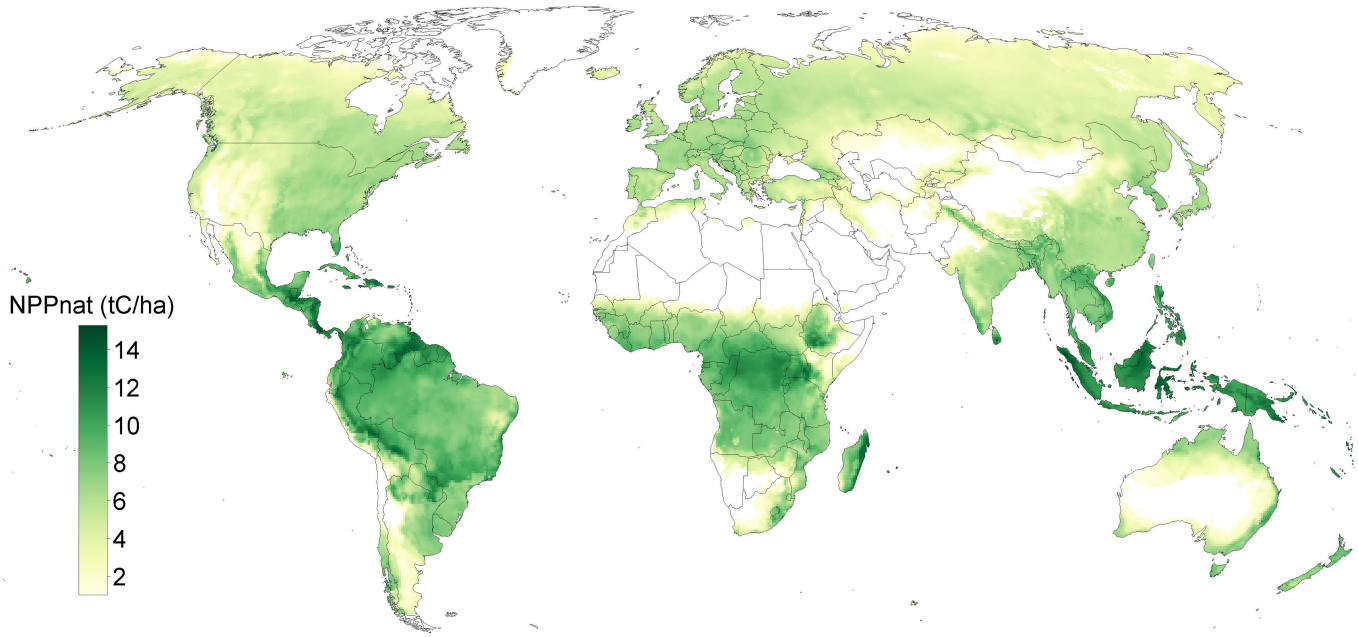
Extended Data Fig. 3 | Above- and below-ground carbon stocks of potential natural vegetation under current climate, used to derive COCs with the carbon loss method. Data simulated with the LPJmL model and adjusted at the biome level according to reference values from the literature (see Supplementary Information).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Soil carbon stocks of potential natural vegetation under current climate used to derive COCs with carbon loss method. Data simulated with LPJmL and adjusted at the biome level according to reference values from the literature (see Supplementary Information).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Annual net primary productivity of potential native vegetation under current climate used to derive COCs with carbon gain method. Data simulated with LPJmL.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Global-average COCs, PEMS and GHGs for a selection of food, feed and fibre items, calculated using the carbon loss method and 4% time discounting

Product category	Product	Carbon opportunity cost ("loss" method)			COC ("gain" method)	Production emissions	TOTAL*			
		Vegetation & mineral soils	Organic soils emissions	Total						
		kg CO2/ kg fresh weight	kg CO2e/ kg fresh weight		kg CO2/ kg fresh weight	kg CO2e/ kg fresh weight	kg CO2e/ kg fresh weight	kgCO2e/ kgDM	kg CO2e/ MJ human kcal†	kg CO2e/ kg protein
Cereals										
	Maize grains	2.0	0.1	2.1	2.3	0.5	2.6	2.9	0.8	29
	Rice grains (rough)	2.4	0.2	2.6	2.2	2.2	4.8	5.5	2.0	69
	Wheat grains	1.8	0.1	1.9	2.5	0.7	2.6	2.9	0.9	23
	Barley grains	2.4	0.2	2.6	3.3	0.5	3.1	3.5	1.0	29
	Sorghum grains	4.0	0.4	4.4	6.8	0.4	4.9	5.5	1.6	55
	Millet grains	3.9	0.9	4.9	7.2	0.5	5.4	6.1	1.8	68
Tubers										
	Cassava tubers	1.5	0.2	1.7	1.3	0.0	1.7	4.9	1.6	160
	White potato tubers	0.6	0.0	0.6	0.6	0.1	0.7	3.4	1.1	38
	Sweet potato tubers	1.2	0.1	1.2	0.9	0.1	1.3	4.5	1.4	90
	Yam tubers	1.4	0.2	1.5	1.2	0.0	1.5	5.1	1.4	100
Sugar crops										
	Sugar cane stems	0.19	0.0	0.2	0.2	0.0	0.2	0.9	0.7	59
	Sugar beet roots	0.17	0.0	0.2	0.2	0.1	0.2	1.0	0.5	29
Oil crops										
	Soybean seeds	5.7	0.2	5.9	5.3	0.3	6.1	6.7	1.5	17
	Oil palm fruit (bunches	1.6	0.5	2.2	1.3	0.1	2.3	4.3	1.2	120
	Canola seeds	5.2	0.6	5.8	5.0	1.0	6.8	7.4	1.6	32
	Sunflower kernels	4.7	0.2	4.9	7.2	0.8	5.6	6.1	1.0	28
	Groundnut pods	5.6	0.4	6.0	6.8	0.3	6.3	6.8	1.7	36
	Coconuts	2.4	0.5	2.8	3.5	0.1	2.9	6.4	2.1	210
Pulses										
	Common beans	13.6	0.6	14.2	15.3	0.6	14.8	16.5	4.4	69
	Chickpeas	3.7	0.0	3.7	9.1	0.5	4.2	4.6	1.2	20
	Cowpeas	10.6	2.5	13.1	18.3	0.5	13.5	15.4	4.2	57
	Pigeon peas	7.5	0.0	7.5	14.2	0.5	8.0	8.9	2.4	37
	Lentils	5.2	0.7	5.9	9.3	0.5	6.3	7.0	2.0	26
Fruits										
	Banana	1.0	0.1	1.1	0.9	0.1	1.2	4.9	2.0	160
	Plantains	2.9	0.2	3.1	2.3	0.1	3.2	9.1	3.8	450
	Other fruit - temperate	0.9	0.0	0.9	1.1	0.2	1.2	6.5	2.7	180
	Other fruit - tropical	0.9	0.1	1.0	0.9	0.1	1.1	6.0	2.5	170
Vegetables										
	Vegetables	0.68	0.0	0.7	0.6	0.2	0.9	11.4	3.7	76

*Includes organic soil emissions.
†To convert to grams CO_2e per megajoule, divide by 4.18.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Global-average COCs, PEMS and GHGs for a selection of food, feed and fibre items, calculated using the carbon loss method and 4% time discounting (continued from Extended Data Table 1)

Product category	Product	Carbon opportunity cost ("loss" method)			COC ("gain" method)	Production emissions	TOTAL*			
		Vegetation & mineral soils	Organic soils emissions	Total						
		kg CO2/ kg fresh weight	kg CO2e/ kg fresh weight		kg CO2/ kg fresh weight	kg CO2e/ kg fresh weight	kg CO2e/ kg fresh weight	kgCO2e/ kgDM	kg CO2e/ MJ human kcal†	kg CO2e/ kg protein
Vegetable oils										
	Soybean oil	10.5	0.3	10.8	9.8	0.8			1.3	
	Palm oil	7.0	2.3	9.3	5.5	1.8			1.2	
	Palm kernel oil	7.0	2.3	9.3	5.5	1.8			1.2	
	Canola oil	8.0	0.9	8.9	7.7	1.7			1.2	
	Sunflower oil	7.3	0.2	7.5	10.9	1.3			1.0	
	Groundnut oil	12.1	0.8	12.9	14.7	0.8			1.5	
	Maize oil	4.7	0.3	5.0	5.5	1.1			0.7	
	Cotton oil	7.2	0.2	7.4	10.2	2.9			1.2	
Sugars										
	Cane white sugar	1.7	0.2	1.9	1.8	0.3			0.6	
	Beet white sugar	0.8	0.1	0.9	1.0	0.3			0.3	
Meat, dairy and eggs										
	Beef and buffalo meat††	135	9.1	143.9	165.3	44.2	188.2	448.0	102.2	1300
	Sheep and goat meat††	174	11.3	185.7	212.8	42.1	227.7	555.5	112.1	1600
	Cow and buffalo milk	5.8	0.4	6.2	7.1	2.3	8.4	67.0	13.1	260
	Sheep and goat milk	19	1.2	19.9	22.8	4.7	24.6	163.9	25.7	490
	Pork^	14	0.8	14.3	15.1	5.5	19.8	48.3	9.4	150
	Poultry meat^	10	0.5	10.7	11.5	3.7	14.4	36.0	8.4	110
	Eggs	10	0.5	10.7	11.4	3.6	14.3	44.6	10.7	130
Livestock feeds										
	Soybean meal	4.8	0.1	4.9	4.5	0.3				11
	Palm kernel meal	3.3	1.1	4.3	2.6	0.8				31
	Canola meal	3.5	0.4	4.0	3.4	0.7				12
	Sunflower meal	3.2	0.1	3.3	4.9	0.6				11
	Groundnut meal	5.9	0.4	6.3	7.2	0.4				15
	Cotton meal	3.2	0.1	3.3	4.6	1.3				11
	DDGS (maize-ethanol)	2.5	0.1	2.7	3.4	0.5				12
	DDGS (wheat-ethanol)	2.5	0.1	2.6	3.3	0.5				12
Other										
	Coffee beans (green)	29	1.7	31.1	24.9	1.2				
	Tea leaves (dried)	15	0.3	14.9	11.0	1.0				
	Cocoa beans (dried)	39	1.9	40.4	34.4	0.7				
	Cotton lint	2.9	0.1	3.0	4.1	1.2				

*Includes organic soil emissions.
†To convert to grams CO_2e per megajoule, divide by 4.18.
††Average, including meat from dairy animals (refers to whole-carcass weight, including bone and fatty tissue).
^Refers to whole-carcass weight, including bone and fatty tissue (see Methods for sources).

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 3 | Consumption GHG costs for a selection of biofuels

Product category	Product	Carbon opportunity cost ("loss" method)			COC ("gain" method)	Production emissions	Net with gasoline/ diesel substitution*		
		Vegetation & mineral soils	Organic soils emissions	Total			CO2 savings	Net GHG balance	
		g CO2/ MJ GE (LHV)	g CO2e/ MJ GE (LHV)	g CO2e/ MJ GE (LHV)	g CO2/ MJ GE (LHV)	g CO2e/ MJ GE (LHV)	g CO2e/ MJ GE (LHV)	g CO2e/ MJ GE (LHV)	kg CO2e/ liter
Bioethanol									
	Maize ethanol	150	8,2	160	160	72	86	140	3,1
	Wheat ethanol	123	8,2	130	180	104	86	150	3,2
	Sugarcane ethanol	93	9,0	100	100	19	86	35	0,7
Biodiesel									
	Soy methylester	287	8,4	300	270	27	88	230	7,8
	Palm oil methylester	192	60	250	150	50	88	220	7,3
	Canola methylester	218	26	240	210	55	88	210	7,0

GE, gross energy; LHV, lower heating value. See Methods for sources.
*For COC data calculated with the carbon loss method.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0679-9

Palaeolithic cave art in Borneo

M. Aubert[1,2,8]*, P. Setiawan[3,8], A. A. Oktaviana[4,8], A. Brumm[2], P. H. Sulistyarto[4], E. W. Saptomo[4], B. Istiawan[5], T. A. Ma'rifat[5], V. N. Wahyuono[5], F. T. Atmoko[5], J.-X. Zhao[6], J. Huntley[1], P. S. C. Taçon[1], D. L. Howard[7] & H. E. A. Brand[7]

Figurative cave paintings from the Indonesian island of Sulawesi date to at least 35,000 years ago (ka) and hand-stencil art from the same region has a minimum date of 40 ka[1]. Here we show that similar rock art was created during essentially the same time period on the adjacent island of Borneo. Uranium-series analysis of calcium carbonate deposits that overlie a large reddish-orange figurative painting of an animal at Lubang Jeriji Saléh—a limestone cave in East Kalimantan, Indonesian Borneo—yielded a minimum date of 40 ka, which to our knowledge is currently the oldest date for figurative artwork from anywhere in the world. In addition, two reddish-orange-coloured hand stencils from the same site each yielded a minimum uranium-series date of 37.2 ka, and a third hand stencil of the same hue has a maximum date of 51.8 ka. We also obtained uranium-series determinations for cave art motifs from Lubang Jeriji Saléh and three other East Kalimantan karst caves, which enable us to constrain the chronology of a distinct younger phase of Pleistocene rock art production in this region. Dark-purple hand stencils, some of which are decorated with intricate motifs, date to about 21–20 ka and a rare Pleistocene depiction of a human figure—also coloured dark purple—has a minimum date of 13.6 ka. Our findings show that cave painting appeared in eastern Borneo between 52 and 40 ka and that a new style of parietal art arose during the Last Glacial Maximum. It is now evident that a major Palaeolithic cave art province existed in the eastern extremity of continental Eurasia and in adjacent Wallacea from at least 40 ka until the Last Glacial Maximum, which has implications for understanding how early rock art traditions emerged, developed and spread in Pleistocene Southeast Asia and further afield.

Since the 1990s, thousands of rock art images have been documented in the karst caves of the Sangkulirang–Mangkalihat Peninsula in East Kalimantan, a province in the Indonesian portion of Borneo[2–11] (P.S., unpublished observations) (Fig. 1). This remote and difficult-to-access region contains 4,200 km^2 of karst outcrops[9,12] that are formed of late Eocene to early Pliocene limestone[13]. The karst terrain is characterized by densely forested mountain chains and towering cliffs that reach heights of several hundred metres. The Sangkulirang–Mangkalihat Peninsula is adjacent to the edge of the Sunda Shelf—a continental shelf that descends to about 2,500 m in depth—and therefore even during low sea-level stands in the Pleistocene, the karsts were situated essentially at the southeastern tip of Eurasia (Fig. 1). Fifty-two rock art sites have been recorded in eight different karst mountain areas between the Berau and East Kutai districts, spanning a distance of about 100 km. The art is often found in remote, high-level caves that contain little other evidence of human habitation. Few sites in the region have been excavated; the oldest published archaeological remains date to 19,761 ± 87 years before present (BP, taken as AD 1950; an uncalibrated accelerator mass spectrometry ^{14}C date on charcoal)[14]. Previous uranium-series (U-series) and ^{14}C dating of a cave drapery that overlies a hand stencil at Lubang Jeriji Saléh suggested a minimum date of about 10 ka for this motif[15] (Supplementary Information).

On the basis of the superimposition of different styles, the rock art of the Sangkulirang–Mangkalihat Peninsula comprises at least three chronologically distinct phases[9]. The oldest style is characterized by large in-filled, reddish-orange-coloured paintings of animals—mainly the Bornean banteng (*Bos javanicus lowi*), a type of wild cattle that is still extant on the island (Extended Data Fig. 1), but also includes what may be now-extinct taxa[16] as well as hand stencils produced using pigment of the same distinctive hue (Extended Data Fig. 1). The second phase is dominated by hand stencils that are dark purple ('mulberry') in colour, which are often clustered into distinct compositions (Extended Data Fig. 1). Many of these stencils are partly in-filled with painted lines, dashes, dots and small abstract signs that possibly represent tattoos or other marks of social identification, and in some instances hand stencils are linked together by painted mulberry lines that form intricate tree-like motifs, which perhaps symbolize kinship connections. Some older reddish-orange hand stencils appear to have been 'retouched' with mulberry paint to create similar in-filled designs and tree-like motifs (Extended Data Fig. 1). This phase also features small, carefully executed mulberry-coloured paintings of anthropomorphs (Extended Data Fig. 2). These elegant, thread-like human figures—henceforth, 'Datu Saman' following the established term for this style[6]—are sometimes shown in small groups, and are usually portrayed with elaborate headdresses and an array of other objects of material culture that includes possible spear throwers. Some figures are depicted in narrative scenes as hunting or pursuing small deer or as engaged in enigmatic social or ritual activities (for example, 'dancing'; Extended Data Fig. 2). The final rock art phase is characterized by anthropomorphs, boats and geometric designs that are usually executed in black pigments (Extended Data Fig. 1). This rock art style is the only one that has thus far been documented elsewhere in Borneo; it is found at other locations in Indonesia and may be associated with the movement of Asian Neolithic farmers into the region from about 4 ka, or more recently[17,18].

To date the earliest beginnings of cave art production in this region of East Kalimantan, and to establish the timing of stylistic changes in the rock art, we undertook a comprehensive programme of U-series dating of calcium carbonate deposits associated with parietal motifs. Over two field seasons, we collected a total of 15 calcium carbonate samples that were associated with 13 motifs at 6 separate cave sites, and which offered the opportunity to provide minimum and/or maximum dates for the images under study. Individual samples were divided into multiple aliquots (between 3 and 7; 65 in total) and the resultant dates are in stratigraphic order, which demonstrates that uranium and thorium are in closed-system conditions (Supplementary Table 1).

The oldest minimum dates are for a large reddish-orange, solid in-fill painting of an animal at Lubang Jeriji Saléh, for which we obtained dates of 40 ka (sample LJS1) and 39.4 ka (sample LJS1A) (Fig. 2, Extended Data Fig. 3, Supplementary Table 2). The image is incomplete and the animal depicted is therefore unclear, but it appears to be a large ungulate that possibly has a spear shaft protruding from its'

[1]PERAHU, Griffith Centre for Social and Cultural Research, Griffith University, Gold Coast, Queensland, Australia. [2]Australian Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Brisbane, Queensland, Australia. [3]Faculty of Visual Art and Design, Bandung Institute of Technology, Bandung, Indonesia. [4]Pusat Penelitian Arkeologi Nasional (Arkenas), Balitbang Kemendikbud, Jakarta, Indonesia. [5]Balai Pelestarian Cagar Budaya Kalimantan Timur, Samarinda, Indonesia. [6]School of Earth and Environmental Sciences, University of Queensland, Brisbane, Queensland, Australia. [7]Australian Synchrotron, Clayton, Victoria, Australia. [8]These authors contributed equally: M. Aubert, P. Setiawan, A. A. Oktaviana. *e-mail: m.aubert@griffith.edu.au

© 2018 Springer Nature Limited. All rights reserved.

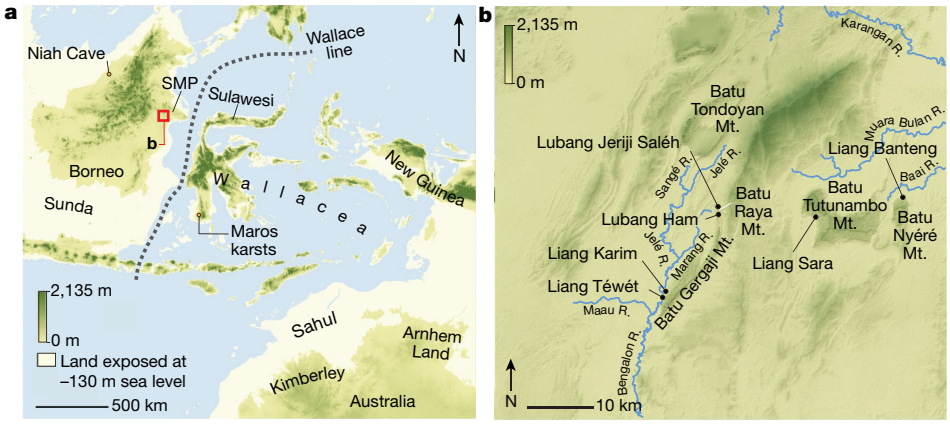


Fig. 1 | Location of the study area. a, Borneo is situated to the west of Sulawesi. The Sangkulirang–Mangkalihat Peninsula is adjacent to the edge of the Sunda Shelf, which is about 2,500 m in depth. Therefore, the karsts were situated at what was essentially the southeastern tip of Eurasia—even during periods of the Pleistocene characterized by low sea levels. **b**, Locations of the archaeological sites included in this study. Map source, Shuttle Radar Topography Mission 1 Arc-Second Global by NASA/NGS/USGS; GEBCO_2014 Grid, version 20150318 (http://gebco.net). Base maps generated using ArcGIS by M. Kottermair and A. Jalandoni.

flank[6]. A figurative painting of a banteng—painted in the same style as the motif that we dated—is located nearby. Additionally, we dated samples LJS5 and LJS6, each of which was associated with a separate reddish-orange hand stencil from Lubang Jeriji Saléh; this provided a minimal date of 37.2 ka for these stencils (Fig. 3). One of these hand stencils—from which sample LJS5 was taken—had previously been dated to a minimum of about 10 ka, but was associated with a large porous cave drapery with suspected open system conditions for uranium and thorium[15] (Supplementary Information). Our new date is associated with dense flowstone underneath the porous cave drapery (Supplementary Information). Another reddish-orange hand stencil at Lubang Jeriji Saléh (sample LJS2) has a minimum date of 23.6 ka and a maximum date of 51.8 ka (Extended Data Fig. 4). In addition, two maximum dates of 103.3 ka (sample LT1) (Extended Data Fig. 5) and 82.6 ka (sample LK1) (Extended Data Fig. 5) were obtained for a reddish-orange hand stencil at Liang Téwét and an animal painting of the same colour at Liang Karim, respectively. These dates correspond to flowstone layers present on the rock face 'canvas' before the art was produced, and thus provide the maximum ages of the images.

The second style of cave art—which, as noted, is characterized by mulberry-coloured paintings— yielded two minimum dates of 16.2 ka (sample LJS4) and 15.7 ka (sample LJS3), both of which were from a single hand stencil from Lubang Jeriji Saléh, as well as a maximum date of 20.9 ka for sample LJS4 (Extended Data Fig. 5). At Liang Banteng, two separate mulberry-coloured hand stencils that feature internal decorations and tree-like motifs with links to other hand stencils were dated to a minimum of 19.7 ka (sample LBT1) (Extended Data Fig. 6) and 17.5 ka (sample LBT2) (Extended Data Fig. 6). At Liang Sara, a mulberry-coloured hand stencil was dated to a minimum of 14.6 ka (sample LSR2) and a nearby Datu Saman figure—also produced using mulberry-coloured pigment—yielded a minimum date of 13.6 ka (sample LSR1) (Fig. 4). This small anthropomorph is depicted wearing a large ornate headdress and brandishing an elongated object, possibly a spear. It is also superimposed over a mulberry-coloured hand stencil. Additional age determinations provided respective minimum dates of 9.3 ka and 0.6 ka for a mulberry-coloured hand stencil (sample LH2) (Extended Data Fig. 7) and an unidentified mulberry-coloured figure (sample LH1) (Extended Data Fig. 8) from Lubang Ham.

In summary, U-series dating shows that the oldest parietal art in eastern Borneo dates to between 51.8 and 40 ka, and that a distinct rock art style appears in the region between 20.9 and 19.7 ka. Concerning the latter, at least some components of the art phase characterized by

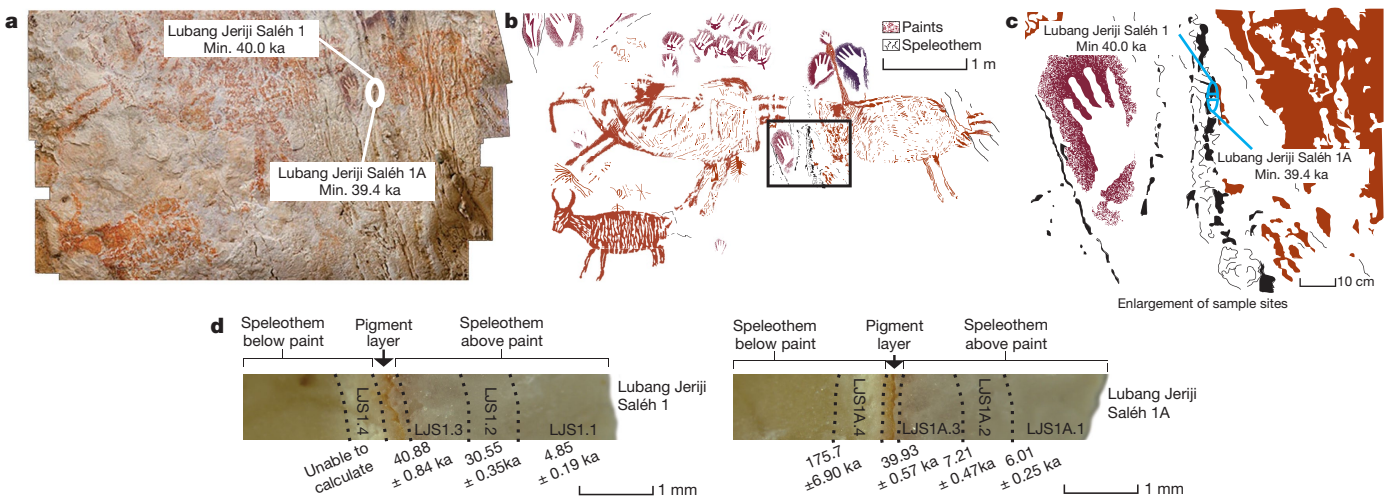


Fig. 2 | Dated rock art from Lubang Jeriji Saléh. Samples LJS1 and LJS1A are shown. **a**–**c**, Photograph (**a**) and tracing (**b**, **c**) showing the locations of the dated speleothems ($n = 2$) and associated painting: a large in-filled reddish-orange-coloured naturalistic depiction of an animal shown in profile. Although the animal figure has deteriorated, we interpret it as a figurative representation of what is possibly a wild bovid (Bornean banteng). Panel **c** is an enlargement of the area marked by a black box in **b**. **d**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. Photograph, L.-H. Fage; tracing, L. Huntley, based on previously published data[6] and at www.kalimanthrope.com.

© 2018 Springer Nature Limited. All rights reserved.

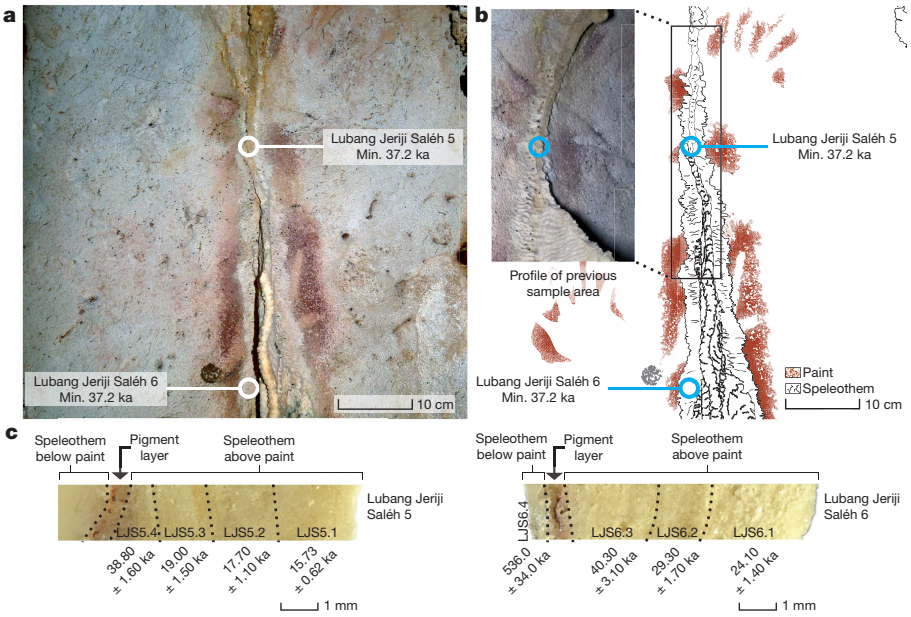


Fig. 3 | Dated rock art from Lubang Jeriji Saléh. Samples LJS5 and LJS6 are shown. **a**, **b**, Photograph (**a**) and tracing (**b**) showing the locations of the dated speleothems ($n = 2$) and associated reddish-orange-coloured hand stencils. A large piece of the cave drapery had previously been removed and dated by a French–Indonesian team[15]. Sample LJS5 is located immediately below the previously sampled cave drapery[6] (shown in **b**) and sample LJS6 is located below the above-mentioned hand stencil and to the left of the cave drapery. **c**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. Tracing, L. Huntley.

mulberry-coloured pigment appear at the height of the Last Glacial Maximum; specifically, hand stencils with internal decorations, and tree-like designs that link the stencils together. The single minimum U-series date that we have on a mulberry-coloured Datu Saman motif suggests that explicit portrayals of human figures had emerged by at least 13.6 ka, although they could potentially have arisen at the Last Glacial Maximum. To our knowledge, the large animal painting from Lubang Jeriji Saléh—created at least 40 ka—is the oldest figurative rock art image in the world. It is also one of the earliest-known figurative representations of an animal, being comparable in age with the mammoth-ivory figurines from the Swabian region of Germany[19,20]. It is also clear that Pleistocene rock art in Sulawesi, where dated motifs ($n = 14$) span the period between about 40 ka and 27.2–22.9 ka[1], is not regionally unique. It is likely that the early parietal art of Sulawesi—which contains hand stencils and paintings of large animals that are stylistically similar to those from Kalimantan (Extended Data Fig. 9)—was introduced from the latter region. Thus, we propose that two major Palaeolithic cave art provinces had emerged at opposite edges of the Eurasian mainland by 40 ka: the renowned Franco-Cantabrian province of western Europe[21] and a province in island Southeast Asia (ISEA) that straddled the Wallace Line.

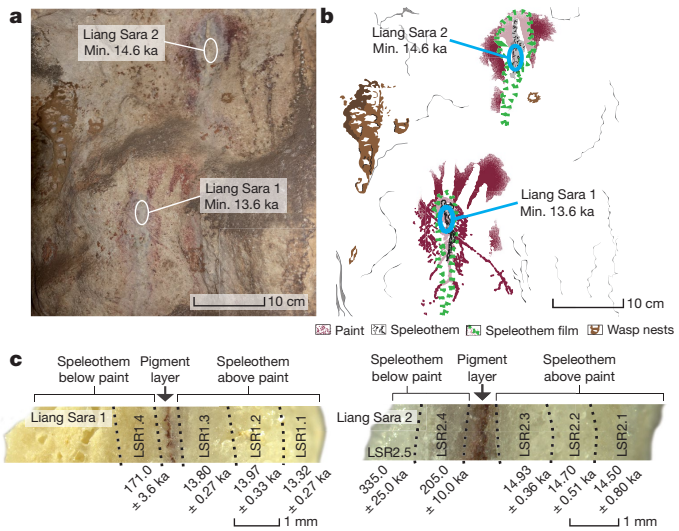


Fig. 4 | Dated rock art from Liang Sara. Samples LSR1 and LSR2 are shown. **a**, **b**, Photograph (**a**) and tracing (**b**) showing the locations of the dated speleothems ($n = 2$) and associated mulberry-coloured hand stencils and human figure. The small anthropomorph is superimposed over a hand stencil in mulberry-coloured pigment. It appears to be wearing a large headdress and holding a spear. **c**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. Tracing, L. Huntley.

It is also apparent that a Pleistocene rock art sequence—which spans at least 20,000 years—arose in Borneo long after the arrival of modern humans in 'Sundaland' (a biogeographical region that encompasses parts of ISEA exposed during times of low sea levels) at around 73–63 ka in Sumatra[22] and long after the peopling of Australia (70–60 ka)[23]. This raises the question of who made the first cave art of Borneo. The oldest anatomically modern human fossil from Borneo—and in ISEA—is the 'Deep Skull', a partial *Homo sapiens* calvarium excavated from the Niah Great Cave (Sarawak, Malaysian Borneo) in the 1950s that has now been dated to about 40 ka (human occupation at Niah Cave dates back to approximately 50 ka)[24]. The Deep Skull is morphologically closer to modern-day eastern Asians than to Australo-Melanesians[25]. Therefore, we suggest two possible models of early modern-human migrations into Pleistocene ISEA. (1) The first *H. sapiens* to enter the region comprised an Australo-Melanesian population that had expanded as far south as Sahul by 70–60 ka but did not produce rock art in ISEA (or, if they did, it has not been discovered or dated); this initial wave was followed by a later migration of an eastern Asian population that arrived about 52–40 ka and produced the earliest rock art in Borneo. (2) Alternatively, the first modern human colonizers in ISEA may have been dispersed into small groups that did not produce rock art, with the latter emerging much later owing to local increases in human population density and flow-on effects on social signalling. Whichever of these models was the case, the absence of evidence for Pleistocene rock art production elsewhere in Borneo reinforces the view that East Kalimantan was an important centre of Palaeolithic cave art development.

© 2018 Springer Nature Limited. All rights reserved.

The distinct rock art style that appears in Borneo at about 21–20 ka is evidence for a major cultural change in Pleistocene ISEA that has not previously been documented. The emergence of this tradition may reflect a population turnover at the Last Glacial Maximum. Alternatively, it is also possible that the karsts of East Kalimantan—one of the richest biodiversity hotspots in ISEA[26]—were a highly favourable environment for human populations. If so, increased levels of inter-group contact in the karsts may have driven the development of a rock art style that was focused on creating visual records of emerging systems of social organization and cultural identity, group affiliation and territorial demarcation. The Datu Saman figures are also notably similar to small anthropomorphs depicted in 'Dynamic Figure' art from Arnhem Land[27] and 'Gwion Gwion'[28] art from the Kimberley (Extended Data Fig. 10), which represent the oldest parietal-art styles from Australia that consistently portray humans; although these styles have long been assumed to be of Pleistocene antiquity, they are not reliably dated. In Franco-Cantabrian cave art, human figures are vastly outnumbered by animal motifs and became common only during the Magdalenian period (16,500–12,000 calibrated years BP)[29]. As noted, with a single minimum date of 13.6 ka, the Datu Saman figures could originate at the Last Glacial Maximum, or they could represent a later stylistic development. In any case, the rock art of the Sangkulirang–Mangkalihat Peninsula documents a clear shift in the development of parietal art from depicting large animals to consistently representing the human world, in the form of human figures and decorated hand stencils with branch-like designs.

It is now evident that rock art emerges in Borneo at around the same time as the earliest forms of artistic expression appear in Europe in association with the arrival of modern humans (45,000–43,000 calibrated years BP)[30]. Thus, similar cave art traditions appear to arise near-contemporaneously in the extreme west and extreme east of Eurasia. Whether this is a coincidence, the result of cultural convergence in widely separated regions, large-scale migrations of a distinct Eurasian population or another cause remains unknown.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0679-9.

Received: 15 June 2018; Accepted: 28 August 2018;
Published online 7 November 2018.

1. Aubert, M. et al. Pleistocene cave art from Sulawesi, Indonesia. *Nature* **514**, 223–227 (2014).
2. Chazine, J. M. Nouvelles perspectives archéologiques à Bornéo, Kalimantan. *Anthropologie* **99**, 667–670 (1995).
3. Chazine, J. M. & Fage, L. H. La ligne de Wallace a-t-elle été franchie par les artistes des temps préhistoriques? Deux nouvelles grottes ornées à Bornéo (E Kalimantan). *Karstologia* **32**, 39–46 (1998).
4. Chazine, J. M. & Fage, L. H. Préhistoire: découverte de grottes ornées à Bornéo. *Archeologia* **352**, 12–19 (1999).
5. Fage, L. H. & Chazine, J. M. *L'art Pariétal des Grottes de Kalimantan (Indonésie). Bilan de 10 Années de Prospection. Découvertes Récentes de Juin 2001 et Perpectives de Protection* (Actes du IIe Congrès national de Spéléologie, Geneva, 2001).
6. Fage, L. H., Chazine, J. M. & Setiawan, P. *Borneo, Memory of the Caves* (Le Kalimanthrope, Caylus, 2010).
7. Setiawan, P. *Monografi Karst Sangkulirang* (Pemerintah Daerah Kutai Timur, Sangatta, 2007).
8. Setiawan, P. *Gambar Cadas Kutai Prasejarah, Kajian Pemenuhan Kebutuhan Terpadu, dan Komunikasi Rupa, Disertasi, Sekolah Pasca Sarjana* (Institut Teknologi Bandung, Bandung, 2010).
9. Setiawan, P. *Gambar Cadas Prasejarah Indonesia* (Direktorat Pelestarian Cagar Budaya dan Permuseuman, Jakarta, 2015).
10. BPCB Kalimantan Timur. *Delineasi Kawasan Sangkulirang Mangkalihat* (Direktorat Pelestarian Cagar Budaya dan Permuseuman, Samarinda, 2016).
11. Fage, L. H. & Chazine, J. M. *Bornéo, la mémoire des grottes* (Fage éditions, Lyon, 2009).
12. Setiawan, P. et al. *Atlas Sangkulirang* (Dinas Lingkungan Hidup, Samarinda, 2012).
13. Wilson, M. E. J., Chambers, J. L. C., Evans, M. J., Moss, S. J. & Nas, D. S. Cenozoic carbonates in Borneo: case studies from northeast Kalimantan. *J. Asian Earth Sci.* **17**, 183–201 (1999).
14. Grenet, M. et al. New insights on the late Pleistocene–Holocene lithic industry in East Kalimantan (Borneo): the contribution of three rock shelter sites in the karstic area of the Mangkalihat peninsula. *Quat. Int.* **416**, 126–150 (2016).
15. Plagnes, V. et al. Cross dating (Th/U-^{14}C) of calcite covering prehistoric paintings in Borneo. *Quat. Res.* **60**, 172–179 (2003).
16. Bellwood, P. *Prasejarah Kepulauan Indo-Malaysia* (PT Gramedia Pustaka Utama, Jakarta, 2000).
17. O'Connor, S. et al. Ideology, ritual performance and its manifestations in the rock art of Timor-Leste and Kisar Island, Island Southeast Asia. *Camb. Archaeol. J.* **28**, 225–241 (2018).
18. Aubert, M. et al. Uranium-series dating rock art in East Timor. *J. Archaeol. Sci.* **34**, 991–996 (2007).
19. Conard, N. J. Palaeolithic ivory sculptures from southwestern Germany and the origins of figurative art. *Nature* **426**, 830–832 (2003).
20. Higham, T. et al. Testing models for the beginnings of the Aurignacian and the advent of figurative art and music: the radiocarbon chronology of Geißenklösterle. *J. Hum. Evol.* **62**, 664–676 (2012).
21. Bahn, P. G. & Vertut, J. *Journey Through the Ice Age* (Weidenfeld & Nicolson, London, 1997).
22. Westaway, K. E. et al. An early modern human presence in Sumatra 73,000–63,000 years ago. *Nature* **548**, 322–325 (2017).
23. Clarkson, C. et al. Human occupation of northern Australia by 65,000 years ago. *Nature* **547**, 306–310 (2017).
24. Barker, G. & Farr, L. (eds) *Archaeological Investigations in the Niah Caves, Sarawak, 1954–2004* (McDonald Institute for Archaeological Research, Cambridge, 2016).
25. Curnoe, D., Datan, I., Taçon, P. S. C., Leh Moi Ung, C. & Sauffi, M. S. Deep Skull from Niah Cave and the Pleistocene peopling of Southeast Asia. *Front. Ecol. Evol.* **4**, 75 (2016).
26. Salas, L. A. et al. Biodiversity, endemism and the conservation of limestone karsts in the Sangkulirang Peninsula, Borneo. *Biodiversity (Nepean)* **6**, 15–23 (2005).
27. Chaloupka, G. *Journey in Time* (Reed Books, Sydney, 1993).
28. Walsh, G. *Bradshaw Art of the Kimberley* (Takarakka Nowan Kas, Toowong, 2000).
29. Fuentes, O. The social dimension of human depiction in Magdalenian rock art (16,500 cal. bp–12,000 cal. bp): the case of the Roc-aux-Sorciers rock-shelter. *Quat. Int.* **430**, 97–113 (2017).
30. Higham, T. European Middle and Upper Palaeolithic radiocarbon dates are often older than they look: problems with previous dates and some remedies. *Antiquity* **85**, 235–249 (2011).

Acknowledgements The fieldwork was authorized by I. M. Geria, the director of the National Centre for Archaeology in Jakarta (Arkenas) and B. Sancoyo and I. M. Kusumajaya, the director and former director of the Balai Pelestarian Cagar Budaya Kalimantan Timur. We further acknowledge the Indonesian State Ministry of Research and Technology for facilitating the research. We thank Griffith University for additional project support. Field assistants included Tewét, Bombé, Amril, Joang, Satriadi, M. Hendra, Stepanus, Satriadi, Sugianor, Heldi, Aidil, Joel, Ghojen, Budiansyah and Firman. Technical laboratory assistance involved A. Nguyen and Y. Feng. We thank S. O'Connor for critical feedback on the manuscript. This research was supported by grants from the Australian Research Council to M.A. (DE140100254 & FT170100025). Part of this work was carried out on the powder diffraction and X-ray fluorescence beamlines at the Australian synchrotron, which is part of ANSTO.

Reviewer information *Nature* thanks M. Bar-Matthews and R. Dennell for their contribution to the peer review of this work.

Author contributions M.A., P.S., A.A.O. and A.B. conceived the study and wrote the manuscript. Site access and project coordination was facilitated by P.H.S., E.W.S., B.I., T.A.M., V.N.W., and F.T.A. Samples were collected and analysed for U-series dating by M.A. and J.-X.Z. Pigment analyses were planned and interpreted by J.H., who also conducted the field emission scanning electron microscopy analysis. Powder diffraction, including phase identification, was conducted by H.E.A.B., and D.L.H. and J.H. conducted the X-ray fluorescence microscopy (XFM). XFM data were analysed and modelled by D.L.H. using GeoPIXE. P.S.C.T. provided comparative analysis between Indonesian and Australian rock art.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0679-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0679-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Dating. A small segment (~100–200 mm^2) of each speleothem was removed from the rock art panels using either a battery-operated rotary tool equipped with a diamond saw blade, or a small hammer and chisel. Each speleothem sample was sawn or chiselled in situ so as to produce a continuous microstratigraphic profile extending from the outer surface of the speleothem through the pigment layer and into the underlying rock face. The only exceptions were LBT1, LBT2 and LH1, in which the sample broke at the speleothem–paint boundary (LBT1 and LBT2) or above the pigment layer (LH1). LT1 and LK1 provided only maximum dates. All of the sampled speleothems comprised multiple layers of dense and non-porous calcite in clear association with painted motifs. In the laboratory, the samples were micro-excavated in arbitrary 'spits' over the surface of the speleothem, creating a series of aliquots ~1 mm thick. The pigment layer was either visible across the entire length of the micro-excavated subsample or was visible at the rear of the samples (LBT1 and BTT2) or on the rock wall (LH1). In total, we obtained 63 U-series age determinations (Supplementary Table 1).

Speleothem samples collected in this study formed from thin films of water on cave surfaces over a long period of time. When precipitated from saturated solutions, calcium carbonate usually contains small amounts of soluble uranium (^{238}U and ^{234}U), which eventually decay to ^{230}Th. The latter is essentially insoluble in cave waters and will not precipitate with the calcium carbonate. This produces disequilibrium in the decay chain, in which all isotopes in the series are no longer decaying at the same rate. Subsequently, ^{238}U and ^{234}U decay to ^{230}Th until secular equilibrium is reached. Because the decay rates are known, the precise measurement of these isotopes enables calculation of the date of the carbonate formation[31].

U–Th dating was carried out using a Nu Plasma multi-collector inductively-coupled plasma mass spectrometer (MC-ICP-MS) in the Radiogenic Isotope Facility at the School of Earth and Environmental Sciences, University of Queensland, following chemical treatment procedures and MC-ICP-MS analytical protocols that have previously been described[32–34]. Powdered sub-samples weighing 3–170 mg were spiked with a mixed ^{229}Th–^{233}U tracer and then completely dissolved in concentrated HNO_3. After digestion, each sample was treated with H_2O_2 to decompose trace amounts of organic matter (if any) and to facilitate complete sample-tracer homogenization. Uranium and thorium were separated using conventional anion-exchange column chemistry, using Bio-Rad AG 1-X8 resin. After stripping off the matrix from the column using double-distilled 7 N HNO_3 as eluent, 3 ml of a 2% HNO_3 solution mixed with trace amount of HF was used to elute both uranium and thorium into a 3.5-ml pre-cleaned test tube, ready for MC-ICP-MS analyses, without the need for further drying down and re-mixing. After column chemistry, the U–Th mixed solution was injected into the MC-ICP-MS through a DSN-100 desolvation nebulizer system with an uptake rate of around 0.1 ml per minute. The U–Th isotopic ratio measurement was performed on the MC-ICP-MS using a detector configuration to enable simultaneous measurements of both uranium and thorium[34,35]. The activity ratios of $^{230}Th/^{238}U$ and $^{234}U/^{238}U$ of the samples were calculated using the previously published decay constants[36]. U–Th dates were calculated using the Isoplot Ex 3.75 Program[37].

It is common for secondary calcium carbonate to be contaminated by detrital materials such as wind-blown or waterborne sediments, a process that can lead to U-series dates that are erroneously older than the true age of the sample. This is due to the pre-existing ^{230}Th present in the detrital components, which is somehow analogous to the radiocarbon marine reservoir effect. As the detrital ^{230}Th cannot be physically separated from the radiogenic ^{230}Th for measurement, its contribution to the calculated ^{230}Th age of the sample is often corrected for by using an assumed activity ratio of $^{230}Th/^{232}Th$ in the detrital component. Given the fact that the detrital component within a cave is often composed of wind-blown or waterborne sediments that chemically approach average continental crust, the mean bulk-Earth or upper continental crustal value of $^{232}Th/^{238}U = 3.8$—corresponding to an activity ratio of $^{230}Th/^{232}Th$ of 0.825—and an arbitrarily assigned uncertainty of 50% have been commonly assumed for detrital or ^{230}Th corrections[32]. In this regard, the degree of detrital contamination may be reflected by the measured activity ratio of $^{230}Th/^{232}Th$ in a sample, with a higher value (such as >20) indicating a relatively small or insignificant effect on the calculated age and a lower value (<20) indicating that the correction on the age will be significant[31]. Because ^{232}Th in the sample is largely present in the detrital fraction and plays no part in the decay chain of uranium, the detrital ^{230}Th in a sample with a measured activity ratio of $^{230}Th/^{232}Th > 20$ would make up only $<0.825/20 = \sim 4.1\%$ of the total ^{230}Th in the sample.

Sometimes the assumed activity ratio of $^{230}Th/^{232}Th$ of 0.825 (±50%) for the detrital component may not cover all situations. If the actual activity ratio of $^{230}Th/^{232}Th$ in the detrital component significantly deviates from this assumed range, the detrital correction scheme may introduce significant bias—especially to samples with an activity ratio of $^{230}Th/^{232}Th < 20$. In such situations, the activity ratio of $^{230}Th/^{232}Th$ in the detrital component can be obtained through direct measurement of sediments associated with speleothems[22,38,39], or computed using isochron methods or stratigraphical constraints[40]. In our case, the speleothem layers associated with painted rock art are very thin and therefore are not suitable for extracting sufficient sediments for direct measurement. On the other hand, these speleothems have extremely low growth rates that are much slower than typical stalagmites used previously[22] and their growths were often episodic, which suggests that the previously published least-squares approach[22] is also not appropriate. Considering the above limitations, we used a limiting stratigraphical constraint method. For instance, using the assumed activity ratio of $^{230}Th/^{232}Th$ of 0.825 (±50%) for the detrital component, the corrected ages of all five sub-samples of LJS2 are in stratigraphic order (Supplementary Table 1), which indicates three episodic growth phases (at 50.3, 26.1 and 16.8–12.1 ka). However, if we increase the assumed activity ratio of $^{230}Th/^{232}Th$ for the detrital component to 1.8, then the corrected age for the stratigraphically younger sample LJS2.2 will be reversed and become older than that for the stratigraphically older sample (see 'Corrected Age-II' in Supplementary Table 1). In this regard, it is reasonable to argue that the activity ratio of $^{230}Th/^{232}Th$ in the detrital component for this cave site should be < 1.8 for the ages to be in stratigraphic order, with the assumed activity ratio of $^{230}Th/^{232}Th$ of 0.825 (±50%) being more reasonable. This is because—using the assumed activity ratio of $^{230}Th/^{232}Th$ of 0.825 (±50%) for detrital correction—the corrected dates of LJS2.1, LJS2.2 and LJS2.3 are 12.6 ± 0.5, 14.9 ± 0.4 and 16.8 ± 1.1 ka, respectively, which are stratigraphically more coherent than other correction schemes (see 'Corrected Age-I' in Supplementary Table 1). We are therefore confident that we used the optimal approach for detrital correction. Regardless, the choice of the correction schemes has almost no effect on the interpretation of the critical samples used to constrain the ages of the rock art. For instance, using the assumed activity ratio of $^{230}Th/^{232}Th$ of 0.825 (±50%) for detrital correction, the corrected dates of LJS1.3, LJS1A.3 and LJS2.5 are 40.9 ± 0.8, 39.9 ± 0.6 and 50.3 ± 1.6, respectively. Using the activity ratio of $^{230}Th/^{232}Th$ of $1.8 \pm 50\%$ for detrital correction, the corrected dates of LJS1.3, LJS1A.3 and LJS2.5 are 39.1 ± 1.8, 38.8 ± 1.2 and 47.2 ± 3.1, respectively, which are indistinguishable within their respective age uncertainties from the former.

Pigment analyses. *Synchrotron powder diffraction.* Small, thin spall flakes of pigment were collected for complementary analyses during sampling for the dating programme. Very small chips of rock art paint (micro-spall) were crushed into homogenized powders manually using an agate mortar and pestle (P1, P2, P3 and P4, Supplementary Information). Sample P1 consisted of mulberry-coloured pigment from the hand stencil located to the left of dating samples LJS1 and LJS1A. P2 consisted of reddish-orange pigment from the animal painting associated with dating samples LJS1 and LJS1A, and P3 consisted of reddish-orange pigment directly related to the dating samples. P4 consisted of mulberry-coloured pigment associated with dating sample LJS4. Once powdered, the rock art paints were placed into 0.3-mm-diameter borosilicate capillaries and mounted on the beam line. Diffraction data were collected at the Australian Synchrotron at a wavelength of 0.77412(3) Å, calibrated using a NIST SRM 660b, from 5-85° 2Theta, with a Mythen microstrip detector with an inherent step size of 0.002°, using two detector positions and a collection time of 5 min per position. Samples were rotated at around 1 Hz during data collection to ensure good powder averaging. Phase identifications were undertaken using Panalytical Highscore with the ICDD PDF4 database.

Synchrotron X-ray fluorescence microscopy. Small, thin spall flakes of pigment were collected for complementary analyses during sampling. We scanned small flakes of rock art paint that adhered to the limestone panel surfaces (samples P1, P2, P3 and P4) at the X-Ray fluorescence microscopy beamline of the Australian Synchrotron using the Maia 384C detector array with incident excitation beam energy of 18.5 keV[41]. The energy resolution of the detector is 275 eV at Mn Kα. The >2-mm-thin chips of paint adhered to limestone were mounted on standard magnetic arms using protective films and adhesive tapes that are invisible to X-rays (chiefly Ultralene). The rock art pigments were positioned at an optimal working distance of 10 mm from the detector. Scans were collected at 2-μm spatial resolution with 1 ms and 2 ms dwell time per pixel for the cross-sections and surfaces, respectively. Scans were collected with full-spectrum X-ray fluorescence data deconvoluted into spatially resolved elemental 'heat' maps using the dynamic analysis method implemented in the GeoPIXE software suite[42].

Scanning electron microscopy. Field emission scanning electron microscopy (JSM-7100F) was used to image surface morphology and the spatial distribution of chemistry was investigated using electron dispersive X-ray spectroscopy (a JED-2300F EDX) undertaken in both spot assay and element-mapping modes. Samples P1 and P2 were platinum-coated for conductivity.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

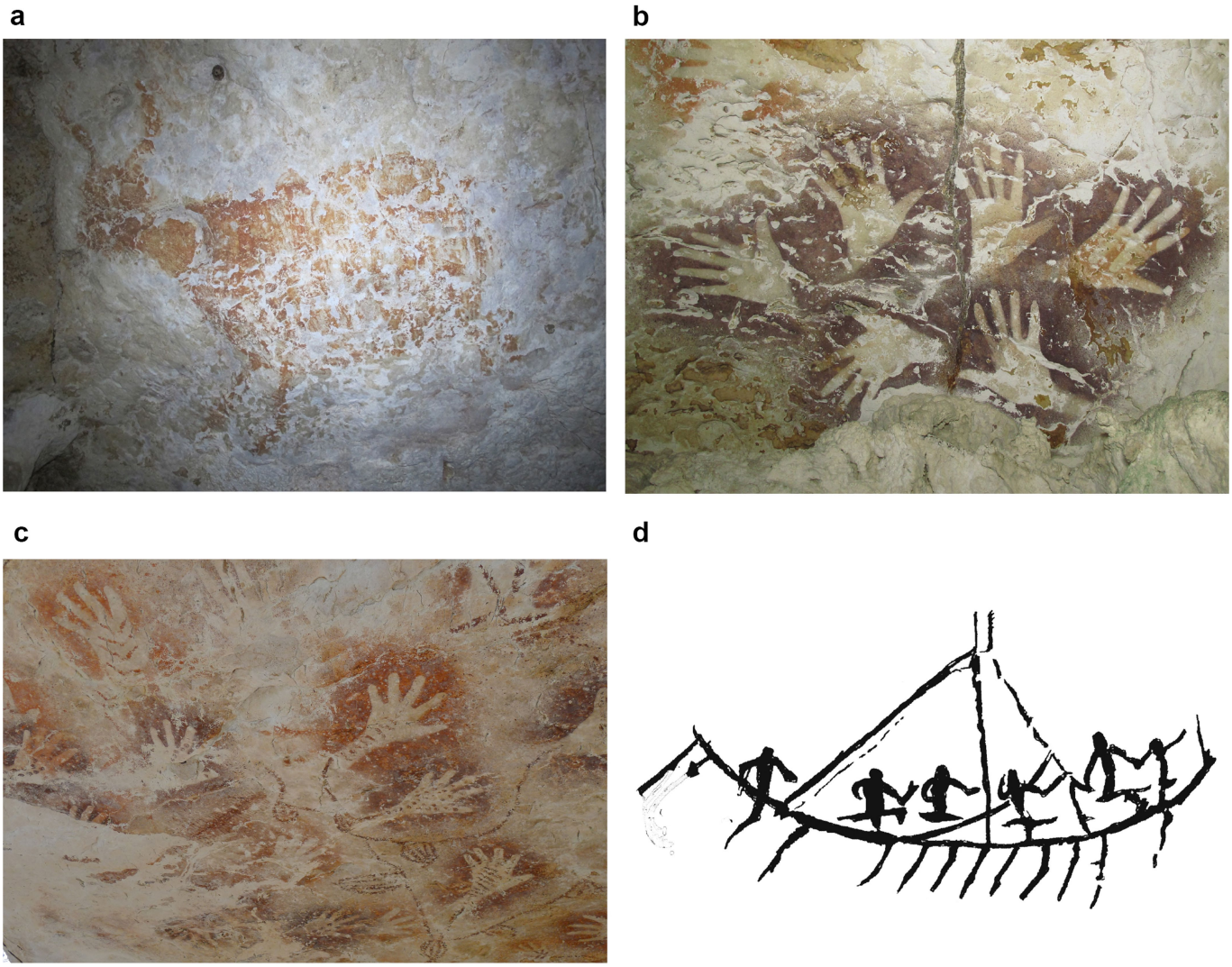
© 2018 Springer Nature Limited. All rights reserved.

Data availability

The authors declare that all the data supporting the findings of this study are available within the paper and its Supplementary Information.

31. Bourdon, B., Henderson, G. M., Lundstrom, C. C. & Turner, S. P. *Uranium-series Geochemistry* (Mineralogical Society of America, Chantilly, 2003).
32. Zhao, J. X., Yu, K. F. & Feng, Y. X. High-precision ^{238}U–^{234}U–^{230}Th disequilibrium dating of the recent past - a review. *Quat. Geochronol.* **4**, 423–433 (2009).
33. Clark, T. R. et al. Spatial variability of initial $^{230}Th/^{232}Th$ in modern *Porites* from the inshore region of the Great Barrier Reef. *Geochim. Cosmochim. Acta* **78**, 99–118 (2012).
34. Clark, T. R. et al. Discerning the timing and cause of historical mortality events in modern *Porites* from the Great Barrier Reef. *Geochim. Cosmochim. Acta* **138**, 57–80 (2014).
35. Zhou, H. Y., Zhao, J. X., Wang, Q., Feng, Y. X. & Tang, J. Speleothem-derived Asian summer monsoon variations in Central China during 54–46 ka. *J. Quat. Sci.* **26**, 781–790 (2011).
36. Cheng, H. et al. The half-lives of uranium-234 and thorium-230. *Chem. Geol.* **169**, 17–33 (2000).
37. Ludwig, K. R. *User's Manual for Isoplot 3.75. A Geochronological Toolkit for Microsoft Excel (Berkeley Geochronology Center Special Publication No. 5)* (Berkeley Geochronology Center, Berkeley, 2012).
38. St Pierre, E., Zhao, J. X. & Reed, E. Expanding the utility of uranium-series dating of speleothems for archaeological and palaeontological applications. *J. Archaeol. Sci.* **36**, 1416–1423 (2009).
39. Hoffmann, D. L. et al. U–Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art. *Science* **359**, 912–915 (2018).
40. Hellstrom, J. U–Th dating of speleothems with high initial ^{230}Th using stratigraphical constraint. *Quat. Geochronol.* **1**, 289–295 (2006).
41. Paterson, D. et al. The X-ray fluorescence microscopy beamline at the Australian Synchrotron. *AIP Conf. Proc.* **1365**, 219–222 (2011).
42. Ryan, C. G. Quantitative trace element imaging using PIXE and the nuclear microprobe. *Int. J. Imaging Syst. Technol.* **11**, 219–230 (2000).

© 2018 Springer Nature Limited. All rights reserved.



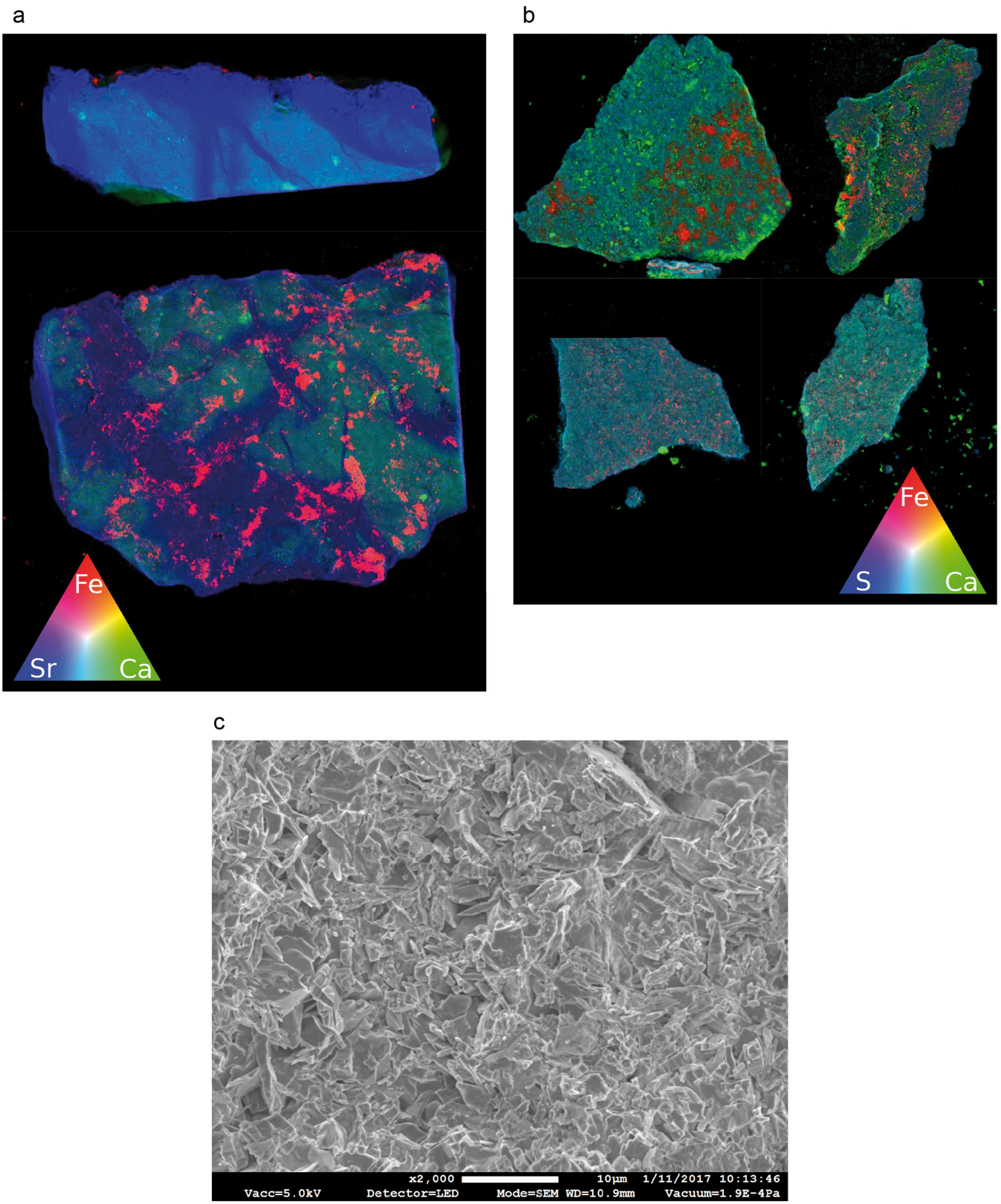
Extended Data Fig. 1 | Rock art styles from the Sangkulirang-Mangkalihat Peninsula. a, The earliest phase of rock art production in the Sangkulirang–Mangkalihat Peninsula is associated with large, in-filled, reddish-orange-coloured paintings of animals and hand stencils. **b**, A second rock art phase is dominated by mulberry-coloured hand stencils—often clustered into distinct compositions and sometimes overlying hand stencils from the previous phase. **c**, Hand stencils from the second phase are often partly in-filled with painted designs and linked together by tree-like motifs, which possibly symbolize kinship connections. Sometimes older reddish-orange hand stencils appear to have been 'retouched' with mulberry-coloured paint and incorporated into these tree-like motifs. **d**, The later rock art phase in the Sangkulirang–Mangkalihat Peninsula is typified by anthropomorphs, boats and geometric designs that are usually executed using black pigments. This style is consistent with early Austronesian iconography, and is possibly related to the arrival of Austronesians in the region at about 4 ka, or more recently.

© 2018 Springer Nature Limited. All rights reserved.



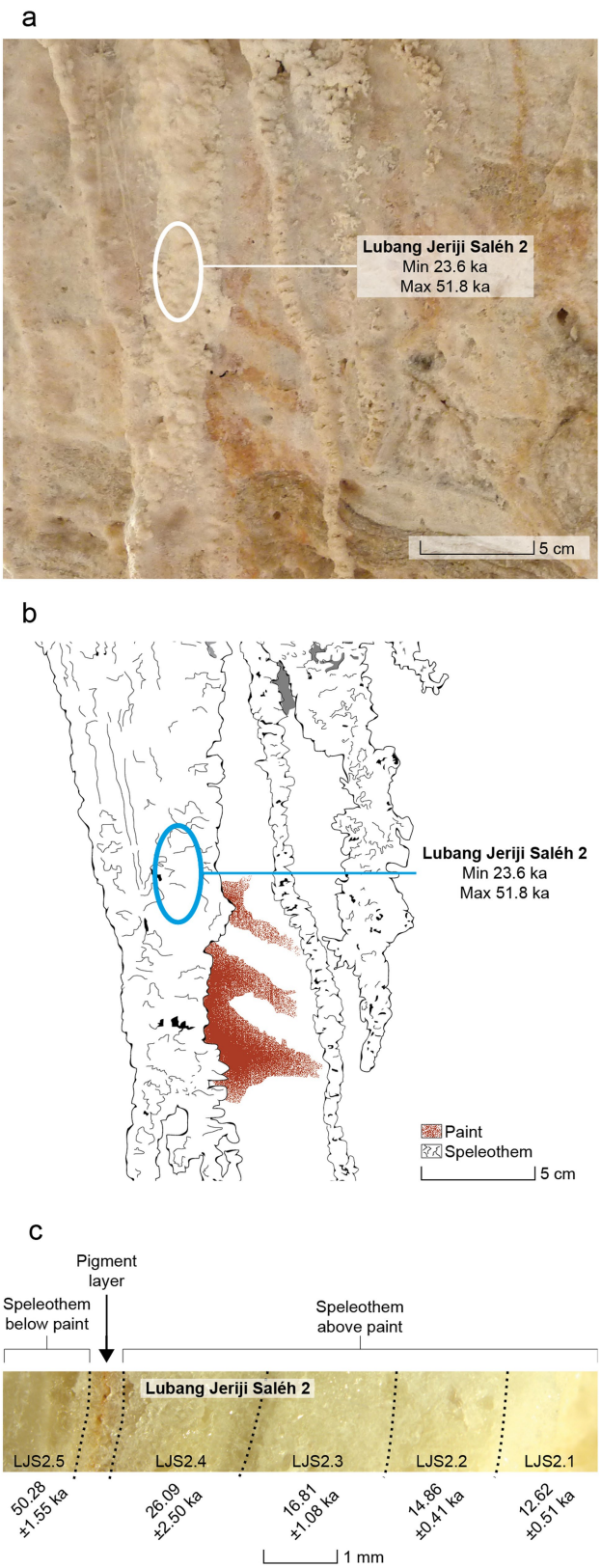
Extended Data Fig. 2 | Example of Datu Saman figures from the Sangkulirang–Mangkalihat Peninsula. The Datu Saman figures of the Sangkulirang–Mangkalihat Peninsula are often depicted in narrative scenes involving small groups with headdresses and other objects.

© 2018 Springer Nature Limited. All rights reserved.



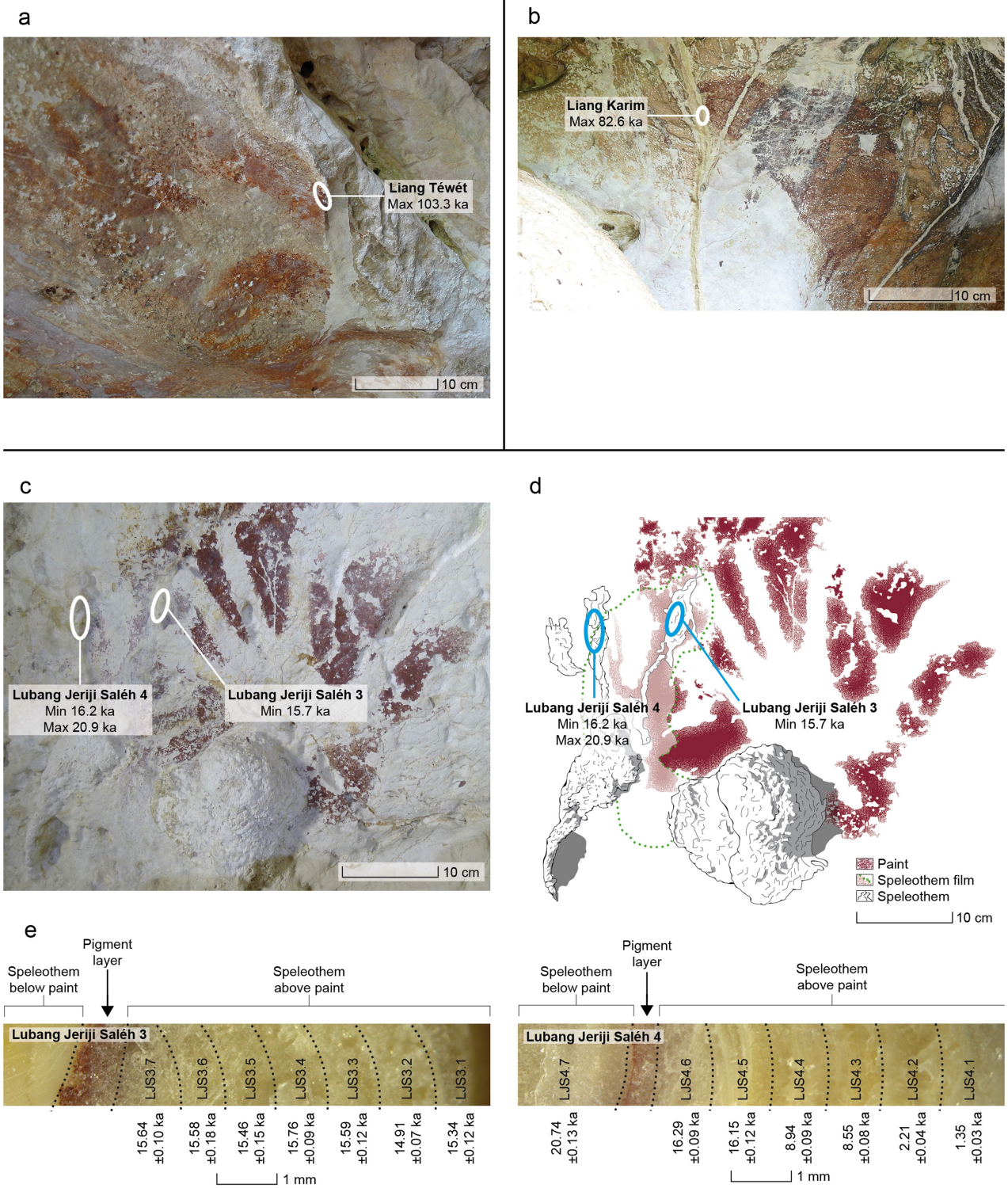
Extended Data Fig. 3 | Pigment analysis. a, P3 cross-section (top) and surface (bottom) showing red–blue–green overlay of XFM element maps for iron, calcium and strontium, respectively. Scans were collected at 2-μm pixel resolution with a dwell time of 1.33 ms per pixel, for a total 2.8 h. **b**, Red–blue–green overlay of XFM element maps for iron, calcium and sulfur for samples P2 (top left), P1 (bottom left), P4a (top right) and P4b (bottom right). Scans were collected at 2-μm pixel resolution with a dwell time of between 1.33 and 2 ms per pixel, for a total 3.5 h. **c**, Scanning electron micrograph of the surface of P1 illustrating typical gypsum crystals overlying larger calcium carbonate grains. Scanning electron microscope data were collected on the P1 sample over five separate scanning electron microscopy sessions between March 2017 and May 2018 with consistent, repeatable results.

© 2018 Springer Nature Limited. All rights reserved.



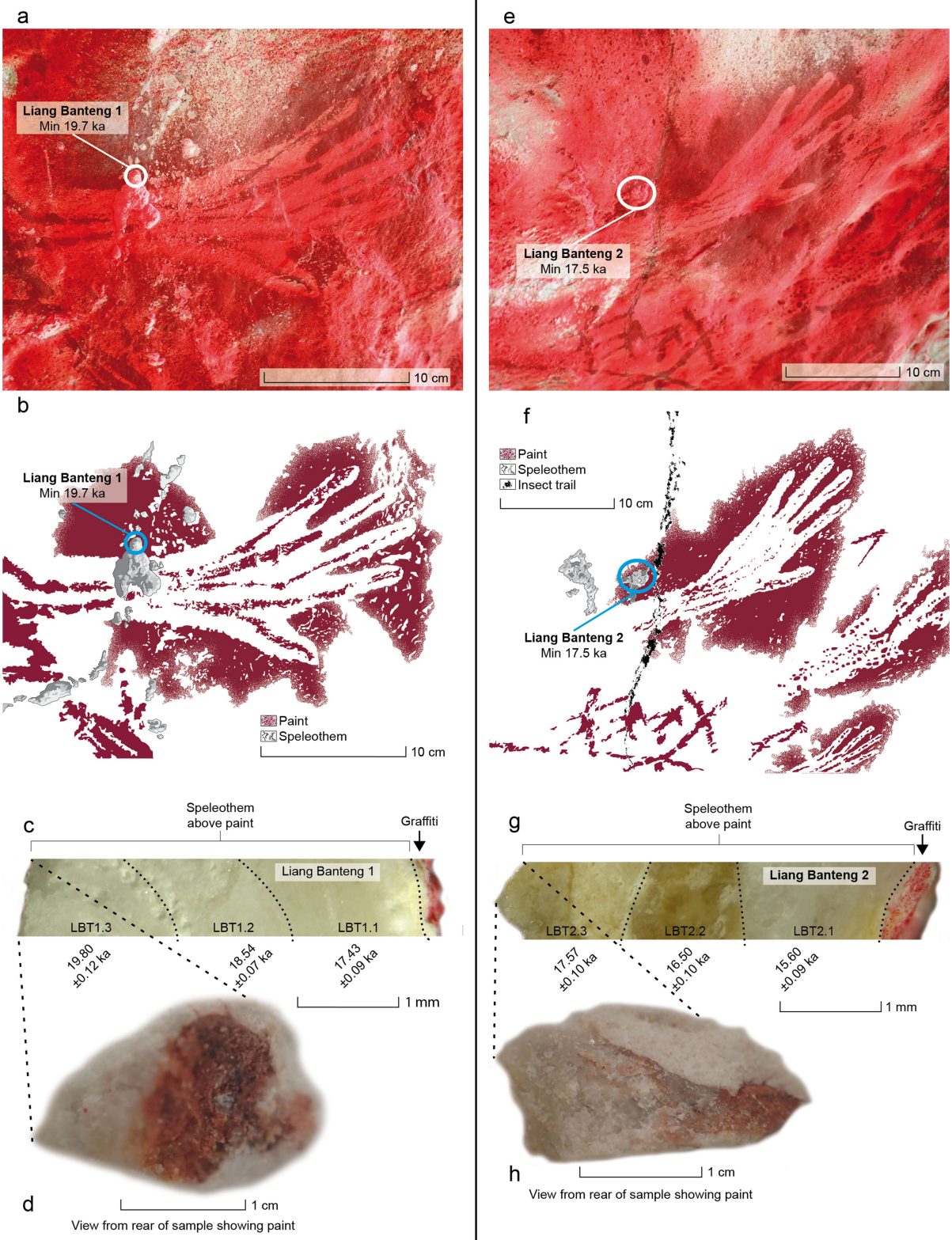
Extended Data Fig. 4 | Dated rock art from Lubang Jeriji Saléh. Sample LJS2 is shown. **a**, **b**, Photograph (**a**) and tracing (**b**) showing the locations of the dated speleothem ($n = 1$) and associated reddish-orange-coloured hand stencil. The date of 51.8 ka provides the maximum date for the earliest rock art phase in the Sangkulirang–Mangkalihat Peninsula. **c**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. Tracing, L. Huntley.

© 2018 Springer Nature Limited. All rights reserved.



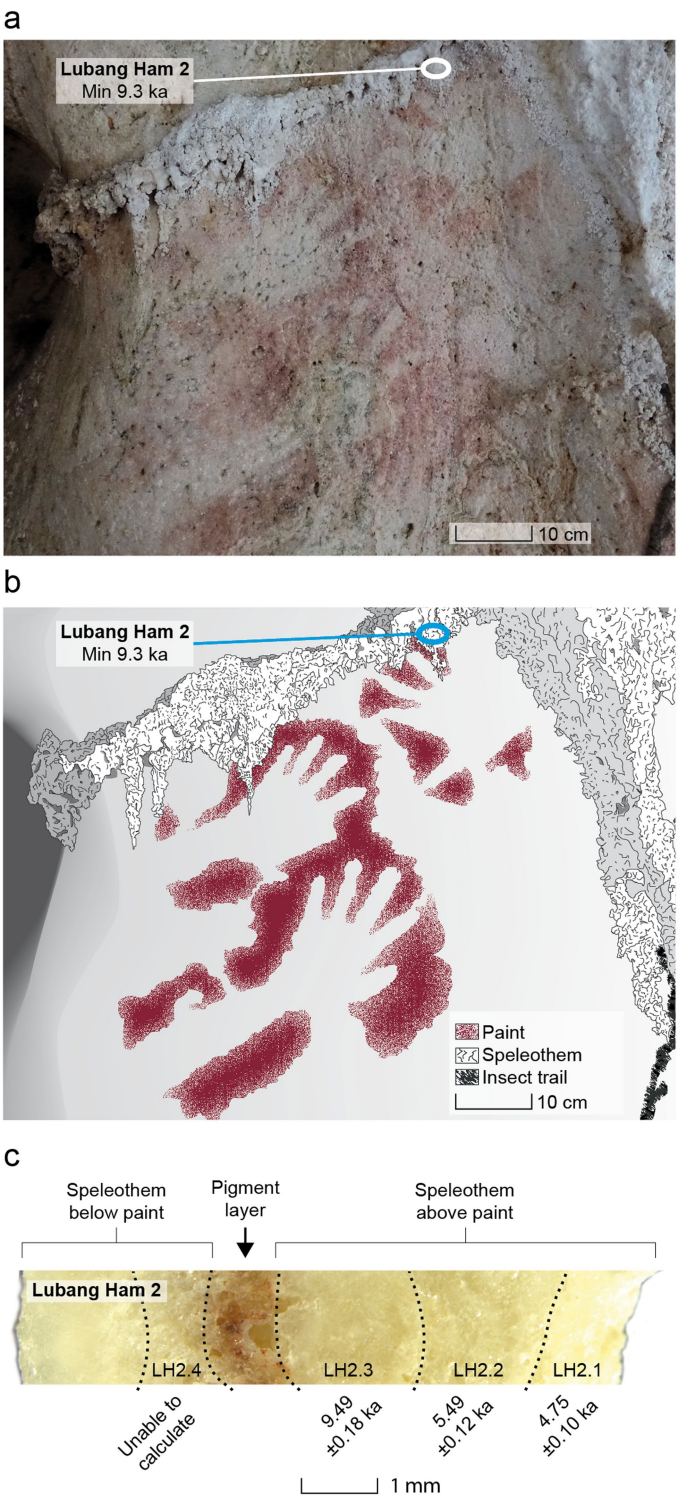
Extended Data Fig. 5 | Dated rock art from Liang Téwét, Liang Karim, and Lubang Jeriji Saléh. Samples LJS3 and LJS4 are shown. **a**, The reddish-orange-coloured hand stencil from Liang Téwét has a maximum date of 103.3 ka. **b**, The animal painting from Liang Karim (possibly a tapir) has a maximum date of 82.6 ka. **c**–**e**, Dated rock art from Lubang Jeriji Saléh (samples LJS3 and LJS4). Photograph (**c**) and tracing (**d**) showing the locations of the dated speleothem ($n = 2$) and associated mulberry-coloured hand stencil. The maximum date of 20.9 ka provides the maximum date for the second rock art phase in the Sangkulirang–Mangkalihat Peninsula. **e**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. Tracing, L. Huntley.

© 2018 Springer Nature Limited. All rights reserved.



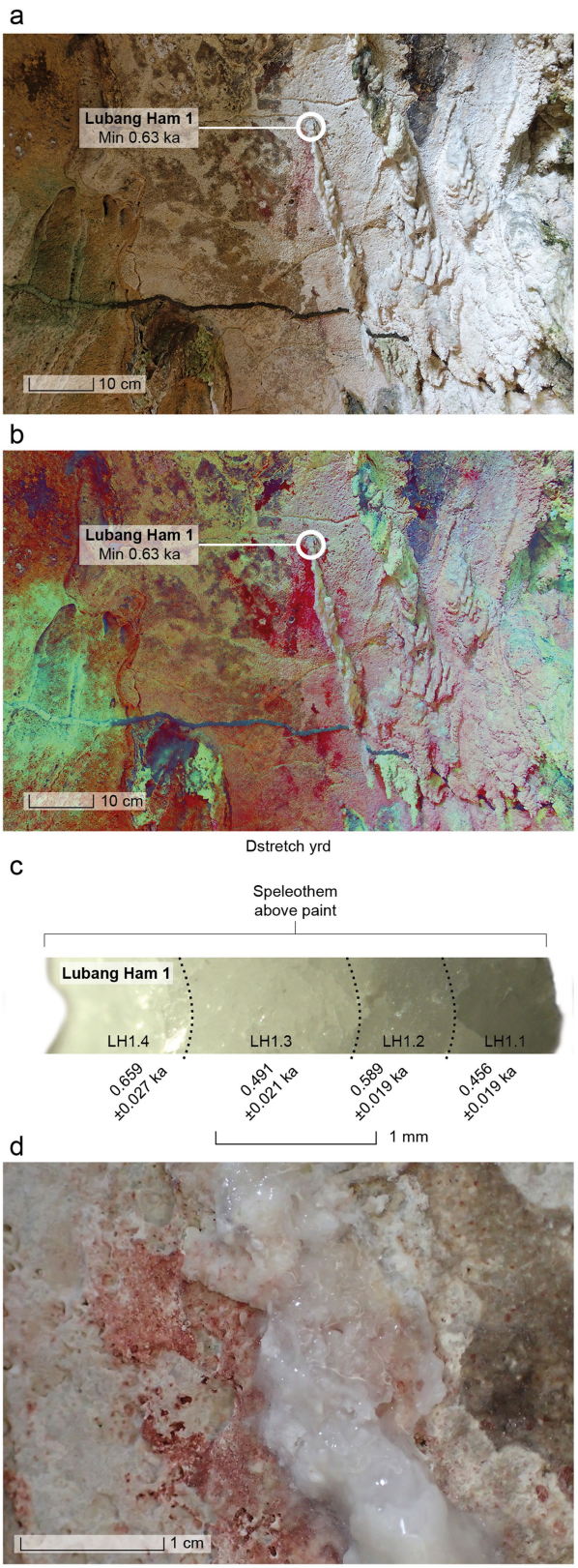
Extended Data Fig. 6 | Dated rock art from Liang Banteng. Samples LBT1 and LBT2 are shown. **a**–**d**, LBT1. **a**, **b**, Photograph (**a**) and tracing (**b**) of sample LBT1 showing the locations of the dated speleothem and associated decorated mulberry-coloured hand stencil. This panel has been the subject of vandalism; it was defaced with bright-red spray paint in 2014 or 2015. **c**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. **d**, The sample broke at the speleothem–paint boundary and the pigment is shown from the rear of the sample. **e**–**h**, LBT2. **e**, **f**, Photograph (**e**) and tracing (**f**) of sample LBT2 showing the locations of the dated speleothem and associated decorated mulberry-coloured hand stencil. This panel has been the subject of vandalism; it was defaced with bright-red spray paint in 2014 or 2015. **g**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. **h**, The sample broke at the speleothem–paint boundary and the pigment is shown from the rear of the sample. Tracing, L. Huntley.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Dated rock art from Lubang Ham. Sample LH2 is shown. **a**, **b**, Photograph (**a**) and tracing (**b**) showing the locations of the dated speleothem and associated mulberry-coloured hand stencil. **c**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. Tracing, L. Huntley.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Dated rock art from Lubang Ham. Sample LH1 is shown. **a**, **b**, Photograph (**a**) and tracing (**b**) showing the locations of the dated speleothem and associated undetermined mulberry-coloured figure. **c**, Profiles of the speleothem showing the micro-excavated subsamples and associated U-series dates. **d**, The sample broke above the pigment layer. Tracing, L. Huntley.

© 2018 Springer Nature Limited. All rights reserved.

a



b



Extended Data Fig. 9 | Large in-filled animal paintings. a, **b**, Large animal paintings from Sangkulirang–Mangkalihat Peninsula (**a**) and south Sulawesi (**b**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Anthropomorph figures from Australia. **a**–**d**, Photographs of rock art from the Kimberley of Western Australia (**a**, **b**) and the Kakadu–Arnhem Land region of Australia's Northern Territory (**c**, **d**). Photographs, M. Donaldson Wildrocks Publication (**a**, **b**) and P.S.C.T. (**c**, **d**).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0740-8

Reward behaviour is regulated by the strength of hippocampus–nucleus accumbens synapses

Tara A. LeGates[1], Mark D. Kvarta[1,2], Jessica R. Tooley[3], T. Chase Francis[4], Mary Kay Lobo[4], Meaghan C. Creed[3] & Scott M. Thompson[1,2]*

Reward drives motivated behaviours and is essential for survival, and therefore there is strong evolutionary pressure to retain contextual information about rewarding stimuli. This drive may be abnormally strong, such as in addiction, or weak, such as in depression, in which anhedonia (loss of pleasure in response to rewarding stimuli) is a prominent symptom. Hippocampal input to the shell of the nucleus accumbens (NAc) is important for driving NAc activity[1,2] and activity-dependent modulation of the strength of this input may contribute to the proper regulation of goal-directed behaviours. However, there have been few robust descriptions of the mechanisms that underlie the induction or expression of long-term potentiation (LTP) at these synapses, and there is, to our knowledge, no evidence about whether such plasticity contributes to reward-related behaviour. Here we show that high-frequency activity induces LTP at hippocampus–NAc synapses in mice via canonical, but dopamine-independent, mechanisms. The induction of LTP at this synapse in vivo drives conditioned place preference, and activity at this synapse is required for conditioned place preference in response to a natural reward. Conversely, chronic stress, which induces anhedonia, decreases the strength of this synapse and impairs LTP, whereas antidepressant treatment is accompanied by a reversal of these stress-induced changes. We conclude that hippocampus–NAc synapses show activity-dependent plasticity and suggest that their strength may be critical for contextual reward behaviour.

Hippocampal activity is altered by changes in the contextual features of rewarding stimuli[3,4], and a population of reward-associated cells

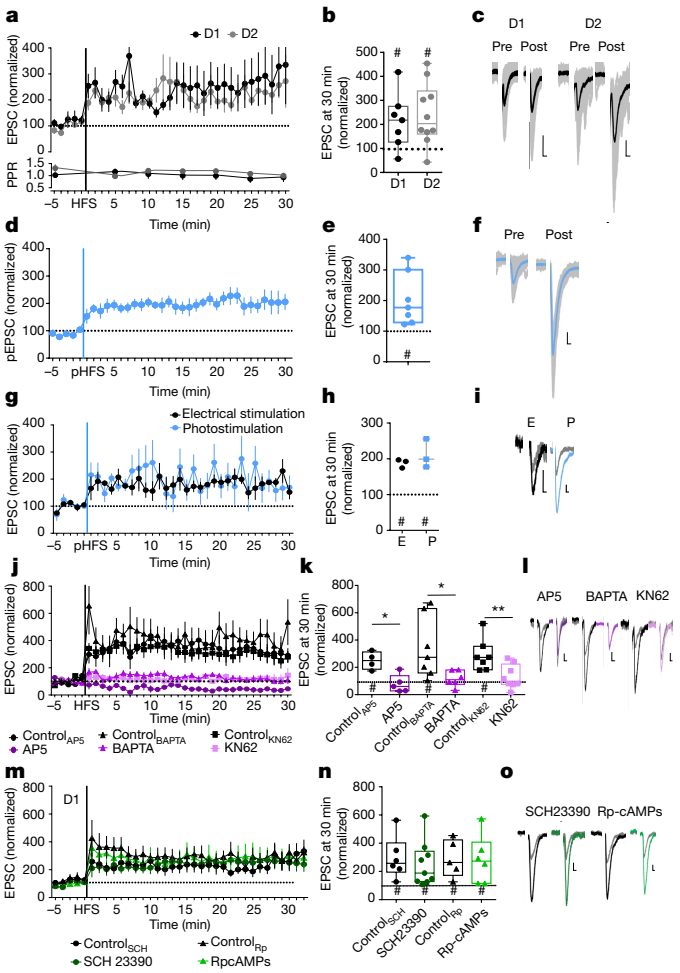


Fig. 1 | Mechanisms that underlie activity-dependent LTP at hippocampus–NAc synapses. a, LTP of hippocampus–NAc eEPSCs is similar in D1R- and D2R-MSNs and does not alter paired pulse ratios. **b**, Summary data from the last 5 min of recording. $^{\#}$D1: $t = 2.624$, $P = 0.0394$, $n = 7$ cells from 7 mice; D2: $t = 3.586$, $P = 0.0059$, $n = 10$ cells from 10 mice. **c**, Representative traces of EPSCs before and after HFS. Grey shading represents individual traces. Black represents the average. **d**, pHFS induces LTP of light-evoked EPSCs. **e**, Summary data from the last 5 min of recording. $^{\#}t = 3.337$, $P = 0.0157$, $n = 7$ cells from 7 mice. **f**, Representative traces of pEPSCs before and after HFS. Grey shading represents individual traces. Blue represents the average. **g**, pHFS potentiates both electrically and optogenetically evoked EPSCs. **h**, Summary data from the last 5 min of recording. $^{\#}$Two-tailed paired Wilcoxon test to compare baseline to response at 30 min: $W = 21$, $P = 0.0313$, $n = 3$ cells from 3 mice. **i**, Representative traces from electrically and optogenetically evoked EPSCs before and after HFS. **j**, Pre-incubation with AP5 or KN62, or chelation of intracellular Ca^{2+} with BAPTA, prevents LTP induction by HFS. **k**, Summary data from the last 5 min of recording. AP5/control$_{AP5}$: $^{*}U = 1$, $P = 0.0317$, $n = 3$, 5 mice; $^{\#}$control$_{AP5}$: $t = 2.865$, $P = 0.0457$, $n = 5$ cells; AP5: $t = 1.729$, $P = 0.1589$, $n = 5$ cells; BAPTA/control$_{BAPTA}$: $^{*}U = 5$, $P = 0.0221$, $n = 7$, 6 mice; $^{\#}$control$_{BAPTA}$: $t = 3.149$, $P = 0.0199$, $n = 7$ cells; BAPTA: $t = 1.172$, $P = 0.2942$, $n = 6$ cells; KN62/control$_{KN62}$: $^{**}U = 6$, $P = 0.0089$, $n = 7$, 8 mice; $^{\#}$control$_{KN62}$: $t = 2.526$, $P = 0.0449$, $n = 7$ cells; KN62: $t = 0.4919$, $P = 0.6378$, $n = 8$ cells. **l**, Representative traces of EPSCs from control cells and cells treated with AP5, BAPTA or KN62. **m**, Pre-incubation with SCH23390 or Rp-cAMPs does not affect LTP induction in D1R-MSNs. **n**, Summary data from the last 5 min of recording. SCH/control$_{SCH}$: $U = 22$, $P = 0.6070$, $n = 6$, 9 mice; $^{\#}$control$_{SCH}$: $t = 5.658$, $P = 0.0013$, $n = 7$ cells; SCH: $t = 2.914$, $P = 0.0195$, $n = 9$ cells; Rp/control$_{Rp}$: $U = 13$, $P = 0.7922$, $n = 5$, 6 mice; $^{\#}$control$_{Rp}$: $t = 2.611$, $P = 0.476$, $n = 6$ cells; Rp: $t = 2.337$, $P = 0.0476$, $n = 9$ cells. **o**, Representative traces of EPSCs from control, SCH23390, and Rp-cAMPs-treated D1R-MSNs. *Differences between treatment and control by two-tailed Mann–Whitney U-test. $^{\#}$Significant increase in EPSC amplitude above baseline revealed by two-tailed paired t-test. LTP kinetics are plotted in 1-min bins. Centre values represent mean, error bars represent s.e.m. For box plots, the middle line is plotted at the median. The box shows the 25th–75th percentiles. Whiskers represent minimum and maximum. Scale bars, 10 pA/10 ms.

[1]Department of Physiology, University of Maryland School of Medicine, Baltimore, MD, USA. [2]Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA. [3]Department of Pharmacology, University of Maryland School of Medicine, Baltimore, MD, USA. [4]Department of Anatomy & Neurobiology, University of Maryland School of Medicine, Baltimore, MD, USA. *e-mail: sthompson@som.umaryland.edu

© 2018 Springer Nature Limited. All rights reserved.

has been identified in this region[5]. Activity-dependent enhancement of spike firing has been observed at hippocampus–NAc synapses[5,6], and cocaine strengthens the connectivity between these two nuclei[7], leading us to hypothesize that plasticity of these excitatory synapses is associated with reward. We first examined whether hippocampus–NAc inputs display activity-dependent synaptic potentiation. Using whole-cell voltage-clamp, we recorded excitatory postsynaptic currents (EPSCs) from medium spiny neurons (MSNs) in the ventromedial NAc shell in brain slices from mice expressing tdTomato in dopamine type 1 receptor (D1R)-expressing MSNs to differentiate between D1R- and presumptive D2R-expressing cells (D1R-MSNs and D2R-MSNs, respectively)[8]. Glutamatergic EPSCs, mediated by both α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA)- and *N*-methyl-D-aspartate (NMDA)-type receptors, were evoked by electrical stimulation of axons of hippocampal cells that projected to the NAc via the fornix. In response to high-frequency stimulation (HFS) (Fig. 1a–c), robust potentiation was elicited similarly in both D1R- and D2R-MSNs. LTP was accompanied by a change in the coefficient of variation and a reduction in transmission failures (Extended Data Fig. 1), but no change in paired-pulse ratio (Fig. 1a), suggesting that postsynaptic expression mechanisms underlie this potentiation.

To verify that fornix-evoked EPSCs were produced by hippocampal input to the NAc, we recorded photostimulation-evoked EPSCs (pEPSCs) in slices from mice expressing channelrhodopsin (ChR) in ventral hippocampus (vHipp) pyramidal neurons (Extended Data Fig. 2). Light pulses delivered in the NAc evoked pEPSCs comparable to EPSCs elicited by electrical stimulation of the fornix. High-frequency photostimulation (pHFS) elicited LTP of a similar magnitude and time course to that elicited by electrical HFS in the fornix (Fig. 1d–f). Furthermore, pHFS potentiated simultaneously recorded EPSCs and pEPSCs evoked with alternating stimuli (Fig. 1g–i).

We next used pharmacological manipulation to dissect the mechanisms that underlie hippocampus–NAc LTP. HFS did not induce LTP in MSNs in the presence of the NMDAR antagonist 2-amino-5-phosphonovaleric acid (AP5), whereas LTP was induced normally in slices in which AP5 was washed out before HFS (Fig. 1j–l). Loading the calcium chelator BAPTA into the cell also blocked induction of LTP, indicating that it is a Ca^{2+}-dependent process (Fig. 1j–l). In accordance with this, LTP induction was blocked by pretreatment of slices with a Ca^{2+}/calmodulin-dependent kinase type II (CaMKII) inhibitor, KN62 (Fig. 1j–l). These properties were observed in both D1R- and D2R-MSNs. Therefore, induction of LTP at hippocampal–NAc synapses requires NMDAR activation, elevation of intracellular [Ca^{2+}], and CaMK activation, much like canonical Schaffer collateral–CA1 cell LTP[9].

An essential mechanism for postsynaptic LTP expression is the insertion of AMPARs. In the ventral tegmental area, Ca^{2+}-permeable AMPARs that lack GluA2 subunits are preferentially inserted during the expression of cocaine-induced plasticity[10]. We investigated whether LTP induction altered subunit composition at hippocampus–NAc synapses. Prior to HFS, hippocampus–NAc EPSCs displayed a linear relationship between current and holding potential, with no change in EPSC amplitude upon application of the selective inhibitor of GluA2-lacking AMPARs, *N*-acetyl-spermine (NASPM) (Extended Data Fig. 3), consistent with the presence of mostly Ca^{2+}-impermeable, GluA2-containing AMPARs at the synapse[11]. Following the induction of LTP, current–voltage relationships remained linear, and EPSCs remained insensitive to NASPM, suggesting that expression of LTP at hippocampus–NAc synapses does not involve insertion of GluA2-lacking AMPARs (Extended Data Fig. 3).

Dopamine is a critical neuromodulator in the NAc, and there is evidence that dopamine signalling is required for LTP induction in the NAc[11–15]. We recapitulated these findings using local stimulation to activate unidentified excitatory synapses within the NAc and found that LTP was blocked in the presence of the D1R antagonist SCH23390 (Extended Data Fig. 4). To examine the requirement of dopamine signalling for LTP specifically at hippocampus–NAc synapses, we recorded

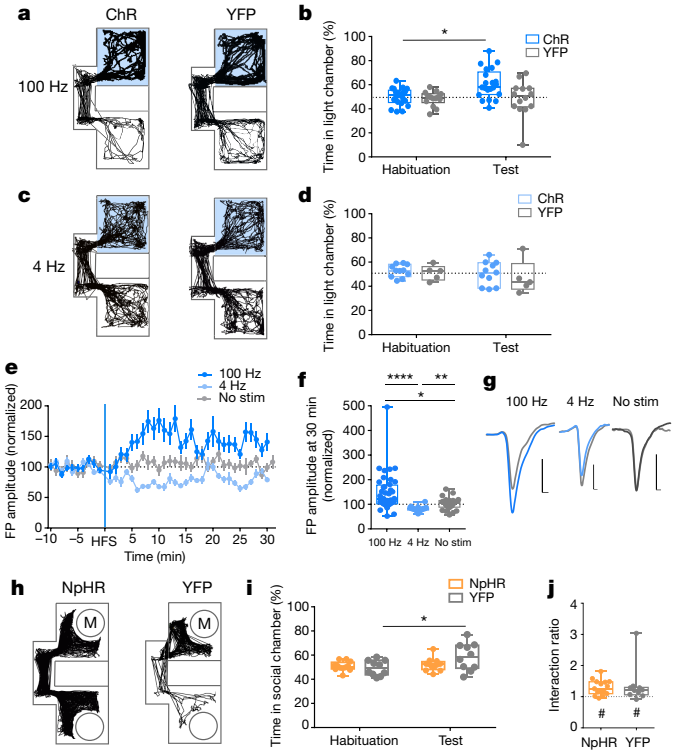


Fig. 2 | In vivo HFS influences reward-related behaviour and NAc activity. a, Representative behavioural trace after 100-Hz conditioning. **b**, Conditioning with 100 Hz induces CPP in ChR-expressing mice. *Two-way repeated-measures ANOVA with Sidak's post hoc test; $F_{1,33} = 5.155$, $P = 0.0298$, $n = 21$, 14 mice. **c**, Representative behavioural trace after 4-Hz conditioning. **d**, Conditioning with 4 Hz light stimulation is not sufficient to induce CPP. Two-way repeated-measures ANOVA: $F_{1,14} = 0.08221$, $P = 0.7785$, $n = 11$, 5 mice. **e**, pHFS induces LTP of vHipp–NAc synapses in vivo. Data are plotted in 1-min bins. Centre values represent mean, error bars represent s.e.m. **f**, Summary data from the last 5 min of recording. Kruskal–Wallis test with Dunn's multiple comparison post hoc test: $H = 34.58$, $P < 0.0001$, $n = 40$, 24, 25 units from 4 mice. **g**, Representative traces of light-evoked LFPs. Scale bars, 0.01 mV/10 ms. **h**, Representative behavioural traces after social interaction conditioning. M: location of the mouse during conditioning. **i**, vHipp–NAc silencing during conditioning blocks social interaction-induced CPP. Two-way repeated-measures ANOVA with Sidak's post hoc test: $F_{1,20} = 4.529$, $P = 0.0459$, $n = 12$, 10 mice. **j**, vHipp–NAc silencing does not disrupt social interaction. Two-tailed Mann–Whitney *U*-test: $U = 64$, $P = 0.5671$, $n = 15$, 10 mice. #One-sample Wilcoxon test shows significant interaction ratios for both groups. NpHR: $W = 114$, $P = 0.0003$, $n = 15$ mice; YFP: $W = 49$, $P = 0.0098$, $n = 10$ mice. For box plots, the middle line is the median. The box represents the 25th–75th percentiles. Whiskers represent minimum and maximum. ****$P < 0.0001$, **$P < 0.01$, *$P < 0.05$.

from D1R- and D2R-MSNs in the presence of their respective receptor antagonists, SCH23390 and sulpiride. Robust LTP was elicited, suggesting that dopamine signalling is not required for the induction of LTP at this synapse in either cell type (Fig. 1m–o; Extended Data Fig. 5). We also examined signalling downstream of dopamine receptors by blocking PKA with Rp-cAMPs, which had no effect on the development of LTP in D1R-MSNs (Fig. 1m–o). Together, these data show that LTP at hippocampus–NAc synapses involves canonical NMDA receptor-dependent mechanisms but does not require dopamine signalling.

To identify a functional role for potentiation at hippocampus–NAc synapses in vivo, we tested whether LTP modulated reward measured by conditioned place preference (CPP). ChR2 or enhanced yellow fluorescent protein (eYFP) were expressed in vHipp. Because collaterals of NAc-projecting hippocampal cells were observed in the prefrontal cortex and amygdala (Extended Data Fig. 6), we implanted fibres into the NAc bilaterally to stimulate hippocampus–NAc synapses selectively. Conditioning of ChR-expressing mice with pHFS resulted in a

© 2018 Springer Nature Limited. All rights reserved.

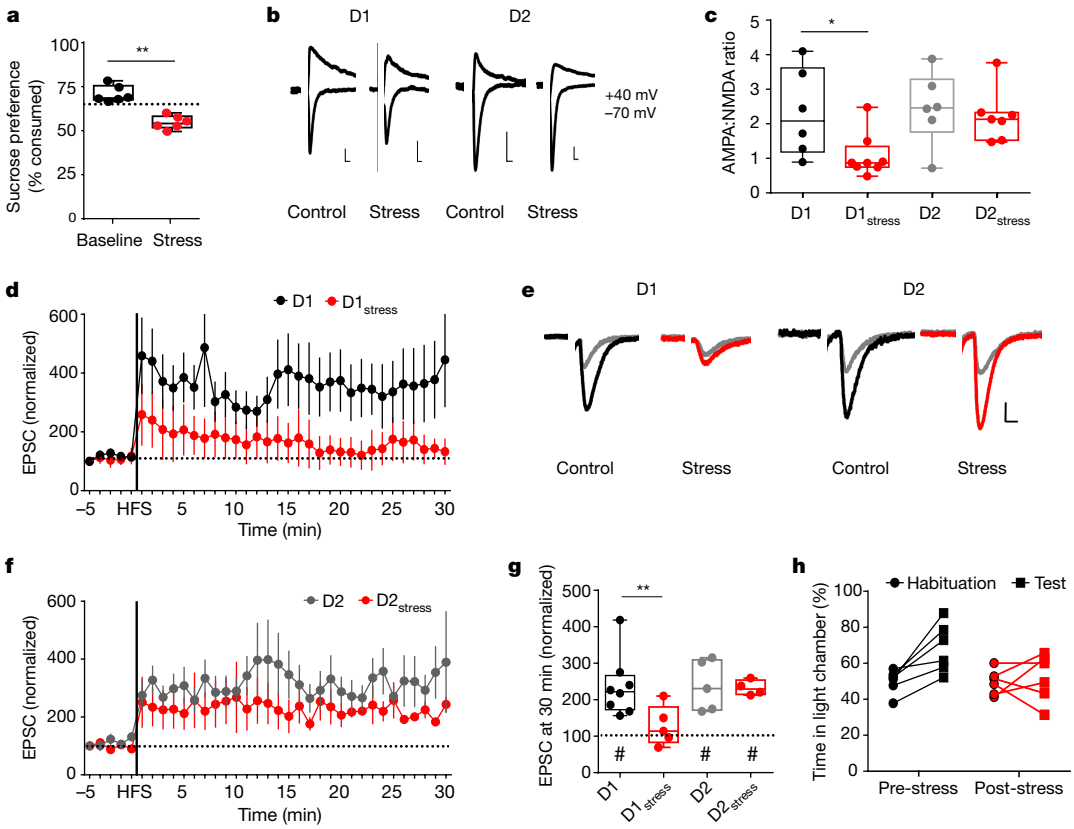


Fig. 3 | Chronic multimodal stress weakens excitatory hippocampal input onto D1R-MSNs. a, Chronic stress induces loss of sucrose preference. Two-tailed paired *t*-test: $t = 5.056$, $P = 0.0039$, $n = 6$ mice. Dotted line represents criterion for anhedonia. **b**, Representative traces of EPSCs at −70 mV and +40 mV from control mice and mice exposed to chronic stress. **c**, Chronic stress decreases AMPA:NMDA ratio. Two-tailed *t*-test: $t = 2.422$, $P = 0.0322$, $n = 6$, 8 mice. **d**, D1R-MSNs from mice exposed to chronic stress show a deficit in LTP induction. **e**, Representative traces of EPSCs from mice exposed to chronic stress and control mice. **f**, Chronic stress has no effect on LTP in D2R-MSNs. **g**, Summary data from the last 5 min of recording. *Two-tailed Mann–Whitney *U*-test: $U = 3$, $P = 0.0109$, $n = 8$, 5 mice; #two-tailed paired *t*-test for baseline EPSC amplitude versus 30 min post-HFS: D1: $t = 3.787$, $P = 0.0068$, $n = 8$ cells; $D1_{stress}$: $t = 1.222$, $P = 0.2564$, $n = 9$ cells; D2: $t = 3.854$, $P = 0.012$, $n = 6$ cells; $D2_{stress}$: $t = 3.164$, $P = 0.0341$, $n = 5$ cells. **h**, Chronic stress abolishes pHFS-induced CPP. Repeated-measures ANOVA with Tukey's post hoc test: $F_{2.109,10.55} = 5.551$, $P = 0.0215$, $n = 6$ mice. LTP kinetics are plotted in 1-min bins. Centre values represent mean, error bars represent s.e.m. For box plots, the middle line is the median. The box represents the 25th–75th percentiles. Whiskers represent minimum and maximum. Scale bar, 10 pA/10 ms.

preference for the light-conditioned chamber, without altering locomotor activity (Fig. 2a, b; Extended Data Fig. 7). eYFP-expressing mice showed no preference for either chamber (Fig. 2a, b). Stimulation at 4 Hz, which does not induce LTP in slices (Extended Data Fig. 8), did not induce CPP (Fig. 2c, d), suggesting that CPP was specifically dependent upon LTP induction.

To demonstrate LTP in vivo, we recorded light-evoked local field potentials (LFPs) in the NAc shell in mice expressing ChR2 in vHipp. We found that HFS induced LTP of light-evoked LFPs (Fig. 2e–g), similar to our whole-cell results. By contrast, LTP was not observed in response to 4 Hz stimulation, or under conditions in which no stimulation paradigm was used (Fig. 2e–g). We also examined optogenetically induced c-Fos expression as a marker of neuronal activation. HFS, but not stimulation at 4 Hz, produced a robust increase in the number of c-Fos$^+$ cells within the NAc shell, but not the core (Extended Data Fig. 9), corresponding to the observed LFP potentiation. We conclude that HFS induces LTP at hippocampus–NAc synapses in vivo, and this presumably underlies the formation of CPP.

We then tested the contribution of this synapse to responses to natural rewards. Mice expressing the light-activated chloride pump halorhodopsin (NpHR) or YFP in the vHipp were tested for CPP in response to social interaction. Light was delivered to the NAc during conditioning to silence activity selectively at hippocampal inputs. YFP-expressing mice displayed a preference for the chamber in which they had previously encountered the target animal, whereas NpHR-expressing mice did not, suggesting that activity of hippocampus–NAc synapses during conditioning is critical for CPP (Fig. 2h, i). By contrast, silencing of hippocampus–NAc synapses did not interfere with the rewarding quality of the social interaction itself. Both NpHR- and YFP-expressing mice showed normal social interaction during light-induced synaptic silencing (Fig. 2j). This suggests that activity at this synapse is not necessary for generalized reward processing, but is necessary for encoding reward associated with spatial context.

Maintaining excitatory drive in the NAc is crucial for normal hedonic state[16,17]. Synaptic weakening in the NAc contributes to stress-induced anhedonia[16], although the source of the input was not identified. We predicted that chronic stress would decrease strength at hippocampus–NAc synapses. We used chronic multimodal stress (CMS) to induce anhedonic-like behaviour, assayed by loss of sucrose preference (Fig. 3a). D1R-MSNs recorded in brain slices taken from mice with a loss of sucrose preference displayed a decrease in synaptic strength, as measured by a decrease in the ratio of AMPAR- to NMDAR-dependent components of the EPSC (AMPA:NMDA ratio; Fig. 3b, c). This is consistent with previous descriptions of stress-induced AMPAR internalization[16]. Furthermore, induction of LTP was profoundly impaired in D1R-MSNs (Fig. 3d, e). By contrast, AMPA:NMDA ratios and LTP were unaltered by chronic stress in D2R-MSNs (Fig. 3b–g). EPSCs in D2R-MSNs instead displayed inward rectification at positive membrane potentials and sensitivity to NASPM (Extended Data Fig. 10), unlike EPSCs in D2R-MSNs in unstressed control mice, suggesting a stress-induced increase in the contribution of Ca^{2+}-permeable, GluA2-lacking synaptic AMPARs. These data demonstrate that chronic stress

© 2018 Springer Nature Limited. All rights reserved.

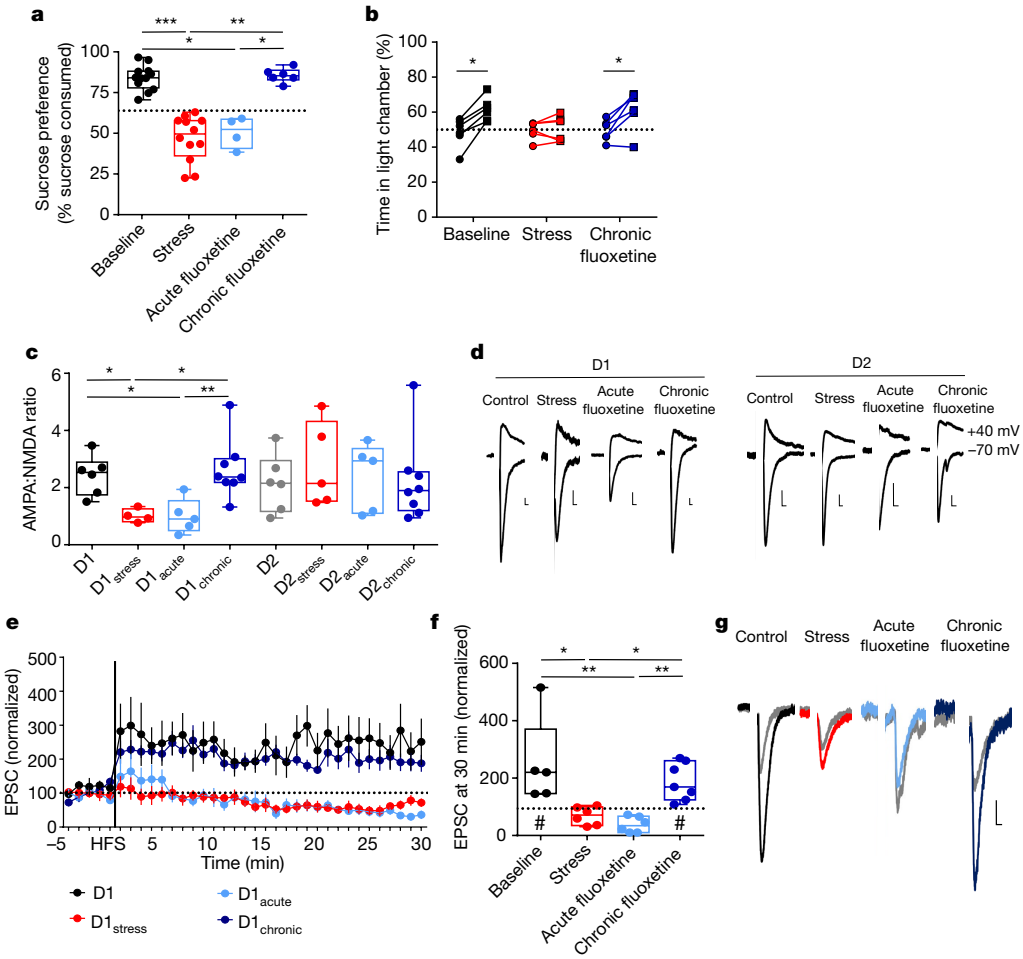


Fig. 4 | Antidepressant treatment rescues synaptic weakening induced by chronic stress. a, Chronic fluoxetine restores normal sucrose preference. One-way ANOVA with Holm–Sidak's post hoc test: $F = 36.38$, $P < 0.0001$, $n = 12, 12, 4, 6$ mice. Dotted line represents anhedonia criterion. **b**, Chronic fluoxetine restores CPP. Two-way repeated-measures ANOVA with Sidak's post hoc test: $F_{2,15} = 7.293$, $P = 0.0061$, $n = 6$ mice. **c**, Chronic fluoxetine restores stress-induced decrease in AMPA:NMDA ratio in D1-MSNs. ANOVA with Holm–Sidak's post hoc test: D1: $F = 7.309$, $P = 0.0019$, $n = 6, 4, 5, 8$ mice. **d**, Representative traces of EPSCs at −70 mV and +40 mV. **e**, Chronic fluoxetine restores LTP deficit induced by chronic stress. **f**, Summary data from the last 5 min of recording. *Kruskal–Wallis test with Dunn's post hoc test: $H = 18.46$, $P = 0.0004$, $n = 5, 8, 6, 7$ mice; #two-tailed paired t-test for baseline EPSC amplitude versus 30 min post-HFS: D1: $t = 4.540$, $P = 0.0027$, $n = 8$ cells; $D1_{stress}$: $t = 0.2615$, $P = 0.8012$, $n = 8$ cells; $D1_{acute}$: $t = 4.109$, $P = 0.0093$, $n = 6$ cells; $D1_{chronic}$: $t = 2.816$, $P = 0.0305$, $n = 7$ cells. **g**, Representative traces of EPSCs before (grey) and after HFS (colour). #Significant increase in EPSC amplitude above baseline revealed by paired t-test. LTP kinetics are plotted in 1-min bins. Centre values represent mean, error bars represent s.e.m. For box plots, the middle line is the median. The box represents the 25th–75th percentiles. Whiskers represent minimum and maximum. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$. Scale bars, 10 pA/10 ms.

selectively weakens the strength and impairs plasticity of hippocampal input to D1R-MSNs. Because activity of D1R-MSNs is associated with positive reward[18–21], these results suggest that the chronic weakening of excitatory drive of the NAc is a contributing factor in stress-induced anhedonia.

As potentiation of the hippocampus–NAc synapse elicits CPP, whereas weakening was associated with anhedonia, we sought to determine the functional consequence of these stress-induced synaptic plasticity deficits. We observed that the ability of pHFS to induce CPP was abolished after exposure to chronic stress, in contrast to results before stress, in which pHFS induced CPP (Fig. 3h). This suggests that chronic stress interferes with CPP by weakening and impairing LTP at hippocampal synapses onto D1R-MSNs.

If dysfunction of hippocampus–NAc synapses contributes to stress-induced changes in reward behaviour, then antidepressant treatment should restore normal reward behaviour and reverse these synaptic changes. We treated mice that displayed loss of sucrose preference after CMS with the selective serotonin reuptake inhibitor fluoxetine. Chronic fluoxetine treatment reversed loss of sucrose preference (Fig. 4a) and restored the CPP deficit induced by chronic stress (Fig. 4b). AMPA:NMDA ratios and LTP in D1R-MSNs from stressed mice treated with chronic fluoxetine were similar to those observed in unstressed controls (Fig. 4c–g). Similarly, stress-induced changes in AMPAR subunit composition observed in D2R-MSNs were restored after chronic fluoxetine treatment (Extended Data Fig. 10). Acute treatment (24–48 h) with fluoxetine, which was not sufficient to restore normal sucrose preference, failed to reverse chronic stress-induced synaptic changes in D1R-MSNs (Fig. 4a–g). Together, these data suggest that restoration of excitatory synaptic strength and plasticity at the hippocampus–NAc synapse coincides with the reinstatement of normal reward behaviour.

Reward drives goal-directed behaviours and various aspects of this process, such as motivation, anticipation, and contextual information, are encoded in different brain regions. We found that synapses formed by hippocampal inputs onto the NAc are highly plastic. Brief correlated high-frequency activity was sufficient to induce both LTP and persistent contextual reward behaviour. Indeed, activity-dependent synaptic plasticity of these synapses is required for the formation of reward-related memories, as shown by the ability of acute silencing of the synapse to disrupt the formation of contextual reward-related memories, but not primary reward processing. Recent work has also shown strengthening of hippocampus–NAc coupling in conjunction with

© 2018 Springer Nature Limited. All rights reserved.

cocaine-induced CPP[7]. The correlation between excitatory strength at this synapse and reward was reinforced by our observation that chronic stress induced deficits in reward-related behaviour, namely anhedonia, and weakened excitatory synaptic strength and impaired plasticity. Conversely, restoration of strength and plasticity at this synapse in response to antidepressant treatment was accompanied by restoration of normal hedonic state. The plasticity of these synapses represents a novel mechanism in the biology of reward. Targeting reward circuits for further study will expand our understanding of the pathophysiology that underlies depression and mechanisms of antidepressant response.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0740-8.

Received: 24 May 2017; Accepted: 1 October 2018;
Published online 26 November 2018.

1. Britt, J. P. et al. Synaptic and behavioral profile of multiple glutamatergic inputs to the nucleus accumbens. *Neuron* **76**, 790–803 (2012).
2. O'Donnell, P. & Grace, A. A. Synaptic interactions among excitatory afferents to nucleus accumbens neurons: hippocampal gating of prefrontal cortical input. *J. Neurosci.* **15**, 3622–3639 (1995).
3. Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D. & Spiers, H. J. Hippocampal place cells construct reward related sequences through unexplored space. *eLife* **4**, e06063 (2015).
4. Tryon, V. L. et al. Hippocampal neural activity reflects the economy of choices during goal-directed navigation. *Hippocampus* **27**, 743–758 (2017).
5. Gauthier, J. L. & Tank, D. W. A dedicated population for reward coding in the hippocampus. *Neuron* **99**, 179–193 (2018).
6. Belujon, P. & Grace, A. A. Critical role of the prefrontal cortex in the regulation of hippocampus–accumbens information flow. *J. Neurosci.* **28**, 9797–9805 (2008).
7. Sjulson, L., Peyrache, A., Cumpelik, A., Cassataro, D. & Buzsáki, G. Cocaine place conditioning strengthens location-specific hippocampal coupling to the nucleus accumbens. *Neuron* **98**, 926–934 (2018).
8. Gerfen, C. R. et al. D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science* **250**, 1429–1432 (1990).
9. Nicoll, R. A. & Malenka, R. C. Contrasting properties of two forms of long-term potentiation in the hippocampus. *Nature* **377**, 115–118 (1995).
10. Terrier, J., Lüscher, C. & Pascoli, V. Cell-type specific insertion of GluA2-lacking AMPARs with cocaine exposure leading to sensitization, cue-induced seeking, and incubation of craving. *Neuropsychopharmacology* **41**, 1779–1789 (2016).
11. Mangiavacchi, S. & Wolf, M. E. D1 dopamine receptor stimulation increases the rate of AMPA receptor insertion onto the surface of cultured nucleus accumbens neurons through a pathway dependent on protein kinase A. *J. Neurochem.* **88**, 1261–1271 (2004).
12. Pawlak, V. & Kerr, J. N. D. Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *J. Neurosci.* **28**, 2435–2446 (2008).
13. Yagishita, S. et al. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616–1620 (2014).
14. Cahill, E. et al. D1R/GluN1 complexes in the striatum integrate dopamine and glutamate signalling to control synaptic plasticity and cocaine-induced responses. *Mol. Psychiatry* **19**, 1295–1304 (2014).
15. Schotanus, S. M. & Chergui, K. Dopamine D1 receptors and group I metabotropic glutamate receptors contribute to the induction of long-term potentiation in the nucleus accumbens. *Neuropharmacology* **54**, 837–844 (2008).
16. Lim, B. K., Huang, K. W., Grueter, B. A., Rothwell, P. E. & Malenka, R. C. Anhedonia requires MC4R-mediated synaptic adaptations in nucleus accumbens. *Nature* **487**, 183–189 (2012).
17. Thompson, S. M. et al. An excitatory synapse hypothesis of depression. *Trends Neurosci.* **38**, 279–294 (2015).
18. Hikida, T., Kimura, K., Wada, N., Funabiki, K. & Nakanishi, S. Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron* **66**, 896–907 (2010).
19. Lobo, M. K. et al. Cell type-specific loss of BDNF signaling mimics optogenetic control of cocaine reward. *Science* **330**, 385–390 (2010).
20. Kravitz, A. V., Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat. Neurosci.* **15**, 816–818 (2012).
21. Pascoli, V., Turiault, M. & Lüscher, C. Reversal of cocaine-evoked synaptic potentiation resets drug-induced adaptive behaviour. *Nature* **481**, 71–75 (2011).

Acknowledgements We thank B. Mathur for D1–tdtomato mice; K. Deisseroth for permission to use the viral optogenetic constructs; the University of Maryland School of Medicine Center for Innovative Biomedical Resources, Confocal Microscopy Core for use of the confocal microscopes; the NIMH Chemical Synthesis and Drug Supply Program for supplying fluoxetine, and T. Gould, T. Blanpied, and T. Bale for suggestions and advice. This work was supported by R01MH086828 (S.M.T.), T32 NS007375 (T.A.L.), a NARSAD Young Investigator Award (T.A.L.), the Whitehall Foundation 2017-12-54 (M.C.C. and J.R.T.), and R01MH106500 (M.K.L. and T.C.F.).

Reviewer information *Nature* thanks J. Dani, C. McClung and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.A.L., M.D.K., T.C.F., M.K.L., and S.M.T. designed experiments. T.A.L. performed viral injections and collected and analysed behavioural and whole-cell electrophysiological data. M.D.K. collected data for behaviour experiments and collected and analysed c-Fos data. J.R.T. and M.C.C. collected and analysed in vivo electrophysiological data. T.C.F. performed viral injections and collected behavioural data. T.A.L. and S.M.T. drafted the article. M.D.K., T.C.F., M.K.L., and M.C.C. provided critical revisions. All authors approved the final version.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0740-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0740-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.M.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Mice. Male DRD1A–tdTomato hemizygous mice were generated by mating a DRD1A–tdTomato hemizygous mouse to a C57BL/6 mouse and were used to differentiate between D1R- and D2R-expressing MSNs. D1R-MSNs were identified by expression of tdTomato whereas unlabelled cells were presumed to be D2R-MSNs. All mice were used between 2 and 4 months of age. Mice were group housed in a 12 h–12 h light–dark cycle with food and water ad libitum. All experiments were performed in accordance with the regulations set forth by the University of Maryland Institutional Animal Care and Use Committee. No statistical methods were used to predetermine sample size. Mice were randomly assigned to experimental and control groups, and experimenters were blinded during data collection and analyses.

Chronic multimodal stress. Sucrose preference was assessed before starting CMS. Only mice that showed a sucrose preference (>70%) were used. Mice were confined to a restraint tube (IBI Scientific, Peosta, IA) in the presence of white noise and a strobe light for 4 h per day, after which they were returned to their home cage and were individually housed. Mice were stressed daily for 10–14 days, and the procedure began no later than zeitgeber time (ZT)4 each day. Loss of sucrose preference (<65%) was used to assess stress susceptibility and defined a depression-like anhedonic state.

Sucrose preference test. Mice were trained by introducing two bottles containing 2% sucrose to their home cage at least one full day before their initial testing. To assess sucrose preference, one bottle containing 1% sucrose in water and one bottle containing plain water were introduced at the beginning of the active (dark) phase. The bottles were removed at the end of the active phase and weighed to measure amount consumed. Sucrose preference was calculated by dividing the volume of sucrose solution consumed by the total volume consumed (water and sucrose) and expressed as a percentage.

Antidepressant treatment. Mice with a sucrose preference (>70%) were subjected to CMS as described above. Upon loss of sucrose preference (<65%), mice were treated with fluoxetine (18 mg/kg/day) in their drinking water acutely (3 days) or chronically (3 weeks). Sucrose preference was tested following fluoxetine treatment.

Electrophysiology. Standard methods were used to prepare 400 μm parasagittal sections that contained both the NAc shell and the fornix, the source of NAc-projecting hippocampal efferents. Dissection and recording were performed in cold artificial cerebrospinal fluid (ACSF) containing (in mM) 120 NaCl, 3 KCl, 1.0 NaH_2PO_4, 1.5 $MgSO_4{\cdot}7H_2O$, 2.5 $CaCl_2$, 25 $NaHCO_3$, and 20 glucose and bubbled with carbogen (95% O_2/5% CO_2). Slices recovered for one hour and were then transferred to a submersion-type recording chamber and superfused at 20–22 °C (flow rate 0.5–1 ml/min).

Cells were visualized under differential interference contrast using a 60× water immersion objective (Nikon Eclipse E600FN). D1R- and D2R-MSNs were identified by the presence or absence of tdTomato, respectively.

Whole-cell currents were recorded in the ventromedial region of the NAc shell under voltage-clamp conditions (–70 mV) using an Axopatch 200B amplifier (Axon Instruments, Molecular Devices) and digitized with a Digidata 1440 analogue-digital converter (Axon Instruments). EPSCs were evoked electrically, by placing a bipolar stimulating electrode (FHC) in the fornix, or optogenetically, by placing a fibre emitting 473 nm blue light from a 473 nm diode-pumped solid-state laser (OEM Laser Systems) above the slice over the NAc shell. EPSCs were evoked at a frequency of 0.1 Hz. Patch pipettes were pulled to resistances of 3–8 MΩ. For LTP experiments, patch pipettes were filled with a solution containing 130 mM K-gluconate, 5 mM KCl, 2 mM $MgCl_6$-H_2O, 10 mM HEPES, 4 mM Mg-ATP, 0.3 mM Na_2-GTP, 10 mM Na_2-phosphocreatine, and 1 mM EGTA. For rectification and AMPA:NMDA ratio experiments, patch pipettes were filled with 135 mM CsCl, 2 mM $MgCl_6$-H_2O, 10 mM HEPES, 4 mM Mg-ATP, 0.3 mM Na_2-GTP, 10 mM Na_2-phosphocreatine, 1 mM EGTA, 5 mM QX-314, and 100 μM spermine. The extracellular solution consisted of ASCF and 50 μM picrotoxin. For experiments involving pharmacological manipulation of signalling pathways involved with LTP induction (AP5 (Sigma-Aldrich, 50 μM), KN-62 (Tocris, 3 μM), Rp-cAMP (Tocris, 5 μM), SCH23390 (Tocris, 3 μM), sulpiride (Tocris, 10 μM)), drugs were superfused over the slice for at least 15 min, after which baseline EPSCs were recorded and HFS was used to elicit LTP. To examine the requirement of Ca^{2+} signalling in LTP induction, BAPTA (Molecular Probes, 10 mM) was included in the patch pipette to block intracellular Ca^{2+}. To examine subunit composition changes after LTP induction, HFS was used to induce potentiation, and NASPM (Tocris, 200 μM) was applied after stable potentiated responses were recorded for 10 min. Recordings were discarded if access resistance changed by >20%.

Summary LTP graphs were generated by averaging the peak amplitudes of individual EPSCs in 5-min bins (six consecutive sweeps) and normalizing these to the mean value of EPSCs collected during the 10-min baseline immediately before the LTP-induction protocol (four bouts of 100 Hz stimulation for 1 s with 15 s between bouts while holding the cell at –40 mV). Individual experiments were then averaged together for graphical representation. The last five minutes of recording were used for statistical comparisons. For AMPA:NMDA ratios, the peak amplitude at –70 mV was used to quantify the AMPA component while the amplitude at +40 mV at 50 ms after stimulation (>3 time constants of the decay of AMPAR-mediated synaptic currents) was used to quantify the NMDA component. The investigator was blind to treatment groups during recording and analysis.

Virus and optogenetic fibre placement surgery. Mice were anaesthetized with 3% isoflurane and underwent stereotaxic surgery to inject serotype 5 adeno-associated viruses (AAV) encoding CaMKIIa-ChR2(H134R) –eYFP, CAMKIIa-eNpHR3.0–YFP, or CaMKIIa–eYFP (UNC Viral Vector Core, Chapel Hill, North Carolina) and implant optic fibres. Virus was injected bilaterally into the vHipp (from bregma anterior/posterior: –3.7, lateral: +3.0, dorsal/ventral: –4.8 from top of skull) and was infused at a rate of 0.1 ml per minute. The injection needle was left in place for 10 min following the infusion. Mice recovered for 6–8 weeks to allow infection of the hippocampal projections to occur. For in vivo optogenetic experiments, 4 mm chronically implantable fibres (0.22 numerical aperture, 105 μm core) were placed bilaterally to target the NAc (anterior/posterior: +1.6, lateral: +1.5, dorsal/ventral: –4.4 from top of skull).

Conditioned place preference. Mice were allowed to recover from surgery for at least two weeks before behaviour experiments. The ability of optogenetic potentiation to induce CPP was evaluated using a three-chamber CPP arena (Maze Engineers), which consisted of two chambers distinguishable by visual cues and a smaller chamber connecting the two rooms. Behaviour was monitored using a camera positioned above the arena, and data were collected using Anymaze software (Stoelting). Mice were allowed to freely explore the entire arena for 30 min. During this habituation phase, mice were connected to a patch cord but no light was transmitted. Mice that showed an inherent preference >65% for either side of the arena were removed from the experiment. On the following day, mice were connected to the patch cord and confined to one compartment during which they were conditioned with ~5 mW 473 nm light administered in four bouts of 100 Hz stimulation for 1 s (2 ms pulse width) with 15 s between bouts using a 473 nm diode-pumped solid-state laser (OEM Laser Systems). The mice remained in the arena for 30 min after stimulation. In a second session on the same day (~4 h later), mice were confined to the other side of the arena while connected to a patch cord with no light administered. Whether mice received light or no light first was randomized. This was repeated and counterbalanced on the following day. Following two days of conditioning, CPP was tested by allowing mice to freely explore the entire arena for 20 min. The experiment was performed similarly for CPP in response to 4 Hz stimulation, except mice were conditioned to light administered in four bouts of 4 Hz stimulation for 25 s with 15 s between bouts.

For the experiments testing the effect of chronic stress and fluoxetine treatment on CPP, mice were subjected to stress, fluoxetine treatment, and CPP as described above. Sucrose preference and CPP were tested in all mice before stress. Mice were then exposed to chronic stress, and once sucrose preference was lost, mice underwent the CPP protocol again. Following this, mice were treated with fluoxetine and restoration of sucrose preference was measured followed by CPP. A separate group of mice continued to undergo stress but were not treated with fluoxetine. The experimenter was blinded to the groups during testing and analysis.

For experiments testing the effect of synaptic silencing, social interaction was used to induce CPP. Set up and habituation were performed as described above. On the following day, mice were connected to the patch cord and confined to one compartment in the presence of a female mouse (target animal) while ~9–10 mW 473 nm light was delivered (3 s on/3 s off) for 30 min. The target mouse was confined in a small wire cage to permit interaction. In a second session on the same day (~4 h later), mice were confined to the other side of the arena while connected to a patch cord with no light administered and in the absence of a target animal. Whether mice received light or no light first was randomized. This was repeated and counterbalanced each day. Following three days of conditioning, CPP was tested by allowing mice to freely explore the entire arena for 20 min.

Social interaction. Social interaction was evaluated in a 33.65 cm × 33.65 cm arena with a 9.5 cm diameter × 10 cm height wire cage positioned on one side to hold the target animal. Behaviour was monitored using a camera positioned above the arena, and data were collected using Anymaze software (Stoelting). Mice were connected to a patch cord to deliver light (~9–10 mW, 473 nm, 3 s on/3 s off) and placed on the side of the arena opposite the target cage in the absence of a target animal. Mice were allowed to freely explore for 150 s after which they were returned to their home cages briefly while a target animal was placed in the target cage. Mice were then placed back into the arena opposite the target cage and allowed to explore for 150 s. The time spent interacting with the target animal was defined by entry into the area immediately surrounding the target cage.

In vivo electrophysiology. CaMKIIa-ChR2(H134R) –eYFP (UNC) was injected into the vHipp as described above and an optic fibre (Thorlabs) was implanted over the vHipp and craniotomy was made over the NAc (from bregma in mm: +1.6 AP, +0.6 ML). A 16-channel, silicone recording probe (A1x16-Poly2-5mm-50 s-177-A16, NeuroNexus) was lowered at a rate of 100 μm/s to a depth of

© 2018 Springer Nature Limited. All rights reserved.

–4.5 mm to target the NAc shell. After allowing 20 min for the recording to stabilize, 10-ms light pulses (473 nm wavelength, Plexbright LED) were delivered through the optic fibre at 2.5 s intervals. After a 10-min baseline recording, 4 stimulation trains (either 100 Hz or 4 Hz, 4 ms pulse width, 15 s interstimulus interval) were applied through the optic fibre as described for slice physiology experiments, before resuming recording conditions identical to baseline for an additional 40 min. A mock stimulation group was included as a control (after 10 min of baseline recording, light pulses were stopped for 2 min before proceeding with recordings). After termination of the recording, the silicone probe was removed, and mice were sutured and returned to their home cage. Each mouse was recorded on each hemisphere for two recording conditions, and the order of hemisphere and stimulation protocol was counter-balanced across subjects. Light evoked-responses in the LFP were analysed using Neuroexplorer software (Plexon). Peri-event histograms of LFP responses were averaged for 24 trials (to yield 1-min intervals) and then computed with 40 ms bins around the onset of light stimulation. The difference between the LFP amplitude in the 40-ms bin immediately preceding light onset and peak LFP deflection was calculated for each 1-min interval and was then plotted as a function of time. Only channels with significant light-evoked changes in the LFP response (as determined by repeated *t*-test during baseline recordings) were used in the analysis.

HFS-induced c-Fos expression. ChR-expressing mice were connected to patch cords while in their home cage, and 100 Hz or 4 Hz blue light was administered as described above. 100 Hz stimulation was administered to YFP-expressing mice. Approximately 70 min later, mice were anaesthetized with isofluorane. Once anaesthetized, the mice were perfused transcardially with 0.9% saline followed by 4% paraformaldehyde. Brains were removed, postfixed overnight in 4% paraformaldehyde, and then transferred to 0.1 M phosphate buffer (PB). Brains were sectioned (40 μm) through the rostro-caudal extent of the NAc using a vibratome. Sections were stored free-floating in 0.1 M PB. Sections were incubated in blocking buffer (0.1 M PB, 3% triton X-100, 0.5% goat serum) for 2 h. Sections were incubated in rabbit anti-c-Fos (Santa Cruz sc-52; 1:1,000) overnight at 4 °C and then visualized with a goat anti-rabbit fluorescent secondary antibody (Alexafluor 546). Sections were mounted on microscope slides and coverslipped with Vectashield. Slides were viewed and imaged on a Nikon Eclipse E400. Photoshop was used to count c-Fos-positive cells and measure the area of region counted from. The number of c-Fos-positive cells was normalized to the area of the region. The investigator was blinded to the groups during processing of tissue and cell counting.

Hippocampus–NAc projection labelling. A retrograde virus expressing Cre recombinase (AAV5-hSyn-Cre-hGH; Penn Vector Core) was injected into the NAc shell (from bregma: anterior/posterior: +1.6, lateral: +0.6, dorsal/ventral: 4.5), and a Cre-dependent virus (AAV2-DIO-ChR2eYFP) was injected into the vHipp (from bregma anterior/posterior: –3.7, lateral: +3.0, dorsal/ventral: –4.8 from top of skull). Viruses were expressed for approximately 8 weeks to allow labelling of hippocampal cells and their projections in the brain. Mice were then perfused as described above and brain postfixed as described above. 100-μm sections were made through the rostra–caudal extent of the brain using a vibratome. Sections were mounted and coverslipped with Vectashield. 10× images were taken using a W-1 spinning disk confocal microscope (Nikon), and *z*-stacks were taken at 100× on a LSM 710 NLO (Zeiss). Maximum intensity projections of the *z*-stacks were generated in ImageJ.

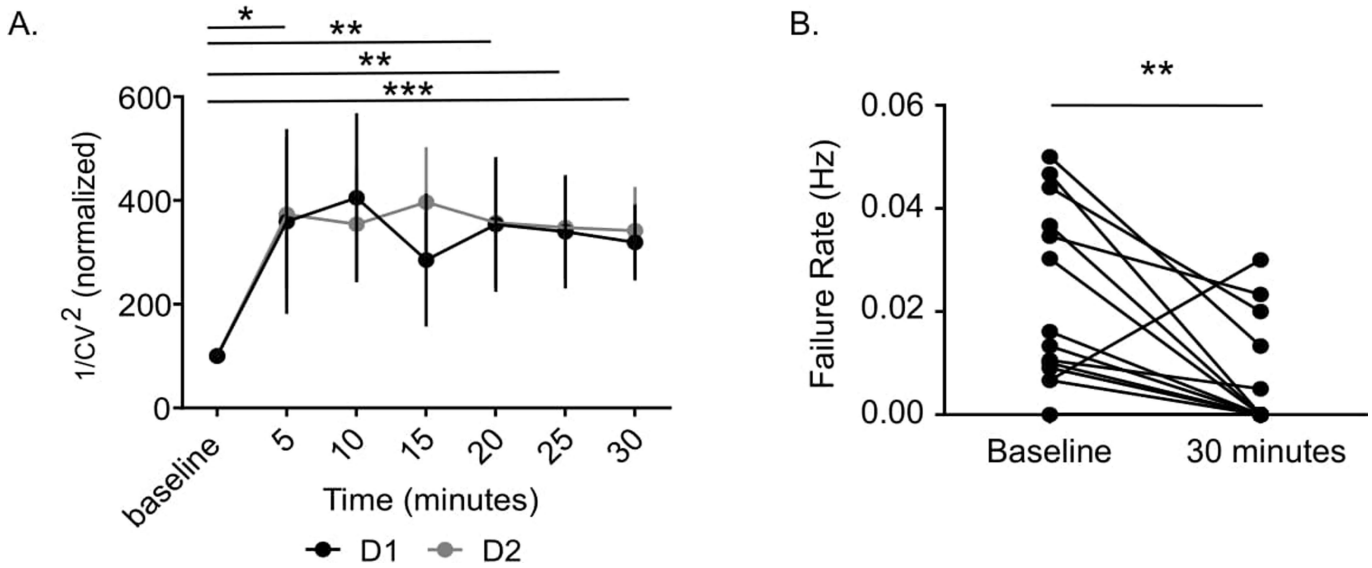
Statistical analyses and data. Statistical analysis was performed using Graphpad Prism 6 software. When results are compared before and after HFS, *n* represents the number of cells or units. For all other experiments, *n* represents the number of mice. For electrophysiological experiments, this represents multiple cells recorded and averaged from each mouse. For box plots, the line in the middle of the box is plotted at the median. The box extends from the 25th to 75th percentiles. Whiskers represent minimum and maximum.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

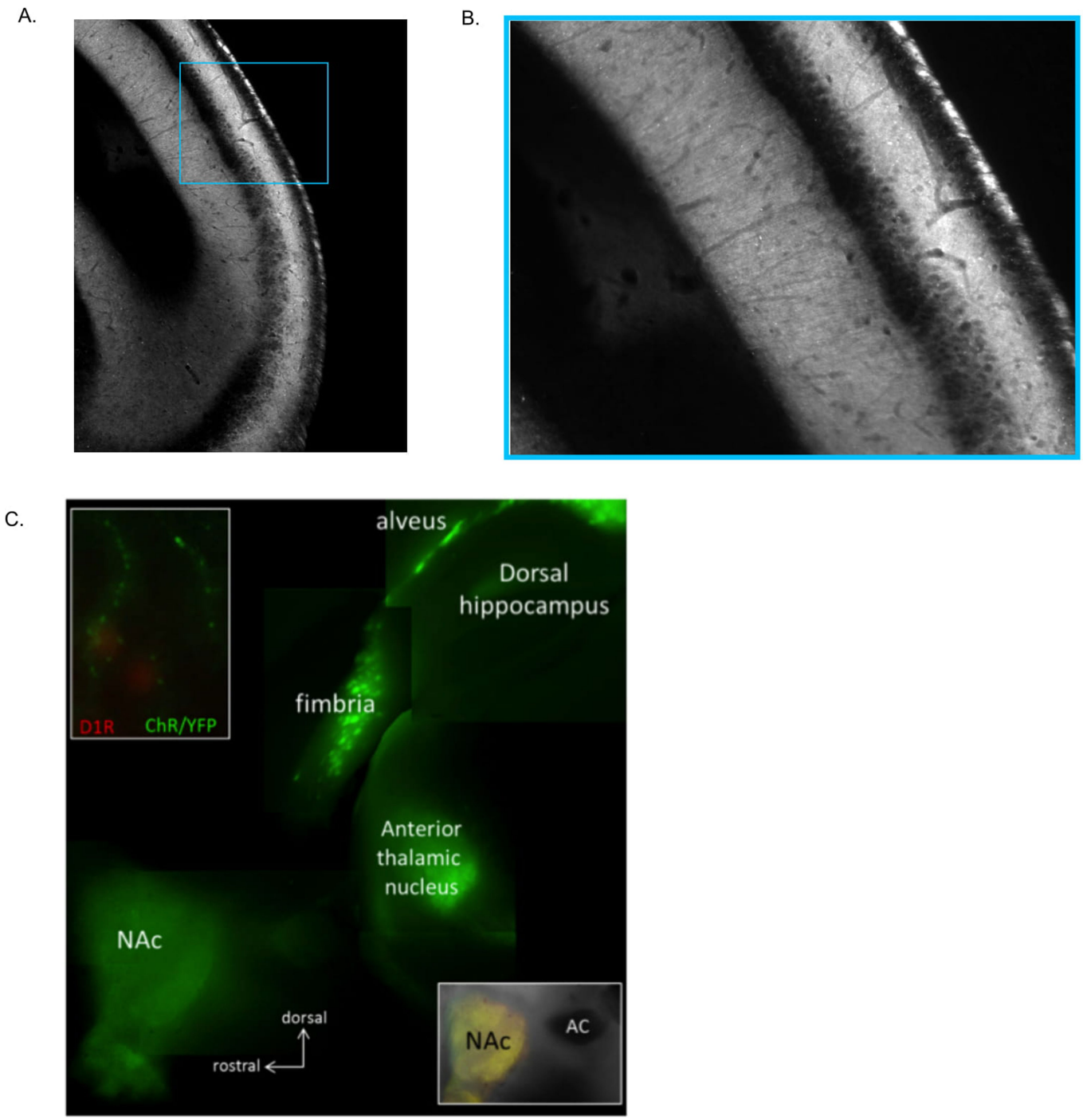
Datasets are available from the corresponding author upon request.

© 2018 Springer Nature Limited. All rights reserved.



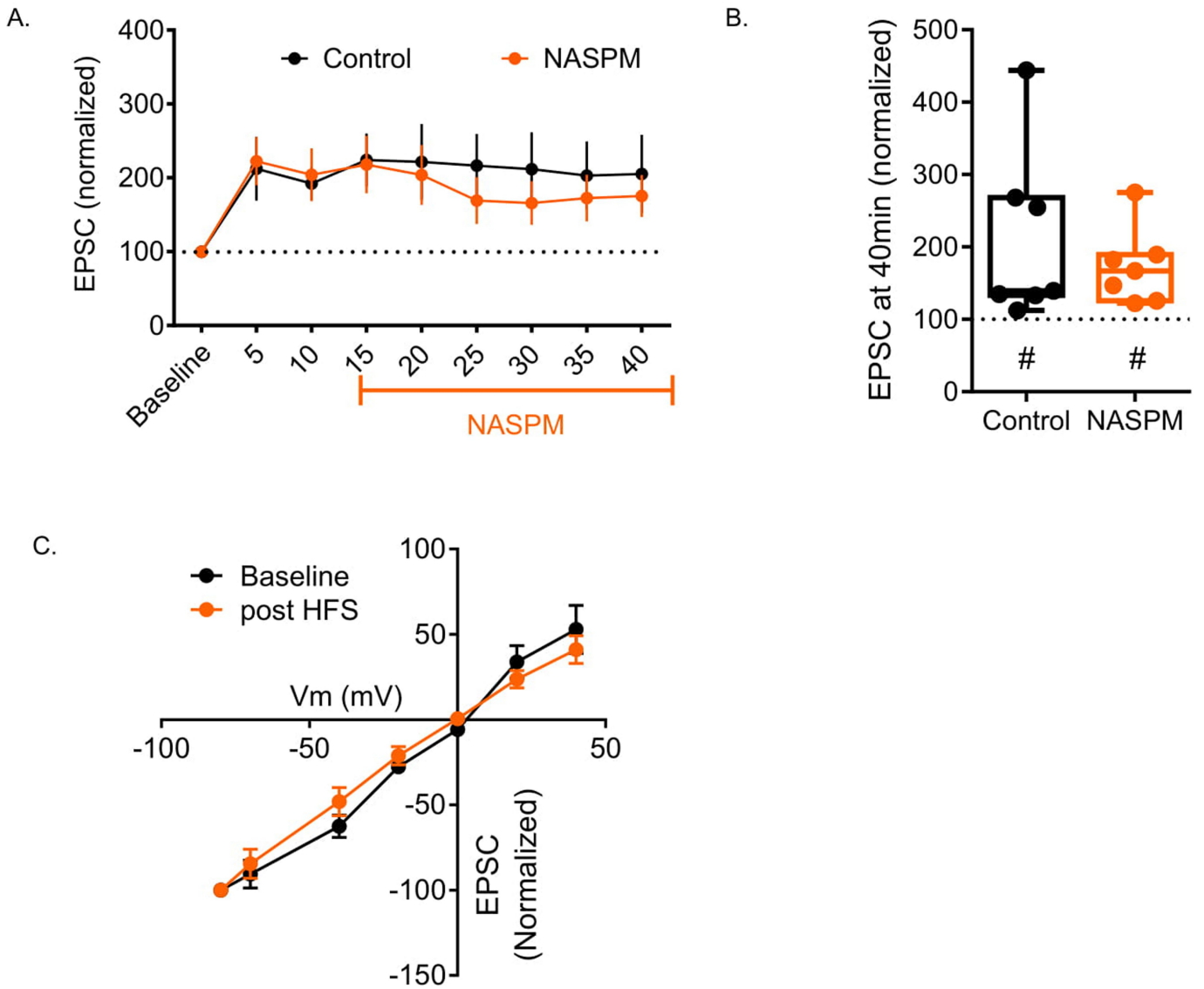
Extended Data Fig. 1 | HFS also induces presynaptic changes and uncovering of silent synapses. a, HFS alters coefficient of variation. Friedman test and Dunn's post hoc test; $Q = 19.95$, $P = 0.0028$, $n = 18$ cells. Centre values represent mean and error bars represent s.e.m. **b**, HFS stimulation decreases failure rate. Two-tailed paired t-test: $t = 3.123$, $P = 0.0066$, $n = 17$ cells. $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$.

© 2018 Springer Nature Limited. All rights reserved.



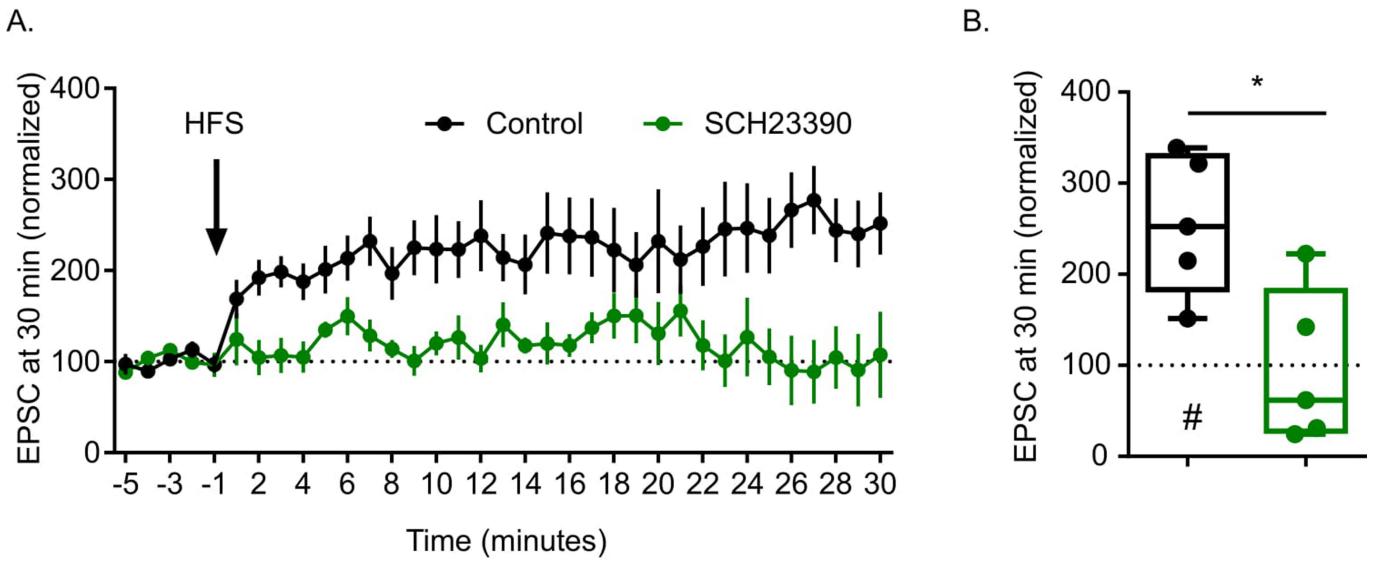
Extended Data Fig. 2 | Representative images of viral injection sites and expression. a, Low-magnification image of YFP fluorescence in ventral hippocampus. **b**, Blue inset from **a**. **c**, Low-magnification image showing YFP fluorescence in the NAc. Insets show overlap in labelling of D1R expression and YFP. This was repeated in mice used for optogenetic experiments. Scale bar, 100 μm.

© 2018 Springer Nature Limited. All rights reserved.



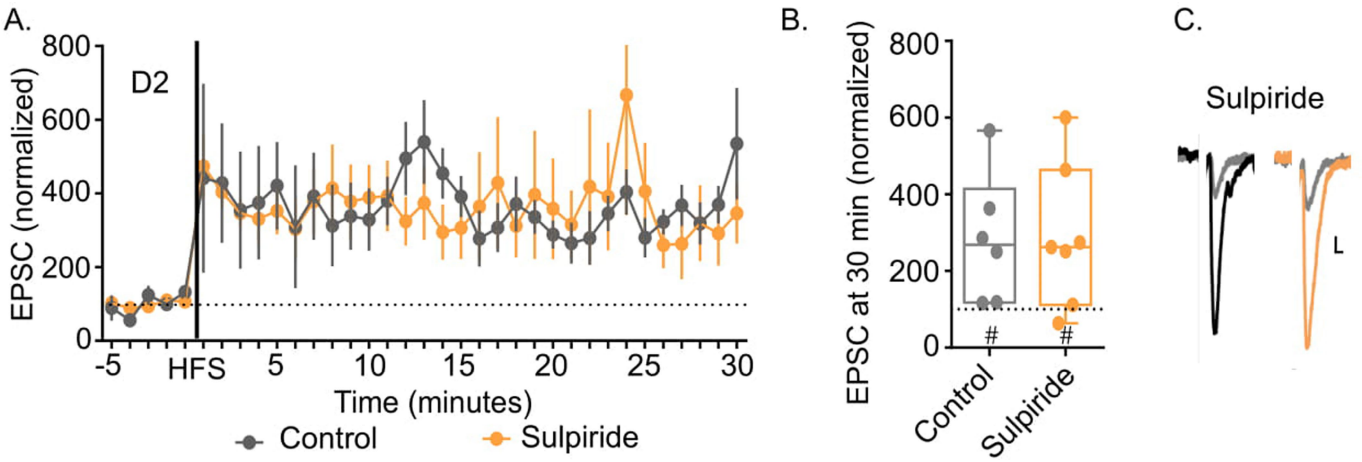
Extended Data Fig. 3 | LTP at hippocampus–NAc synapses is not associated with preferential insertion of GluA2-lacking AMPA receptors. a, Wash-in of NASPM after LTP induction does not alter EPSC amplitude. Centre values represent mean and error bars represent s.e.m. **b**, Summary data from the last 5 min of recording. Two-tailed Mann–Whitney *U*-test: $U = 24$, $P > 0.9999$, $n = 7, 7$ mice; two-tailed paired *t*-test baseline EPSC amplitude/30 min post HFS: control: $t = 2.508$, $P = 0.046$, $n = 7$ cells; NASPM: $t = 2.747$, $P = 0.0226$, $n = 10$ cells. **c**, HFS does not alter rectification at positive holding potentials. Two-tailed Mann–Whitney *U*-test: $U = 12$, $P = 0.7879$, $n = 7$, 4 mice. Centre values represent mean and error bars represent s.e.m. $^{\#}$Significant increase in EPSC amplitude above baseline revealed by paired *t*-test. For box plots, the line in the middle of the box is plotted at the median. The box extends from the 25th to the 75th percentiles. Whiskers represent minimum and maximum.

© 2018 Springer Nature Limited. All rights reserved.



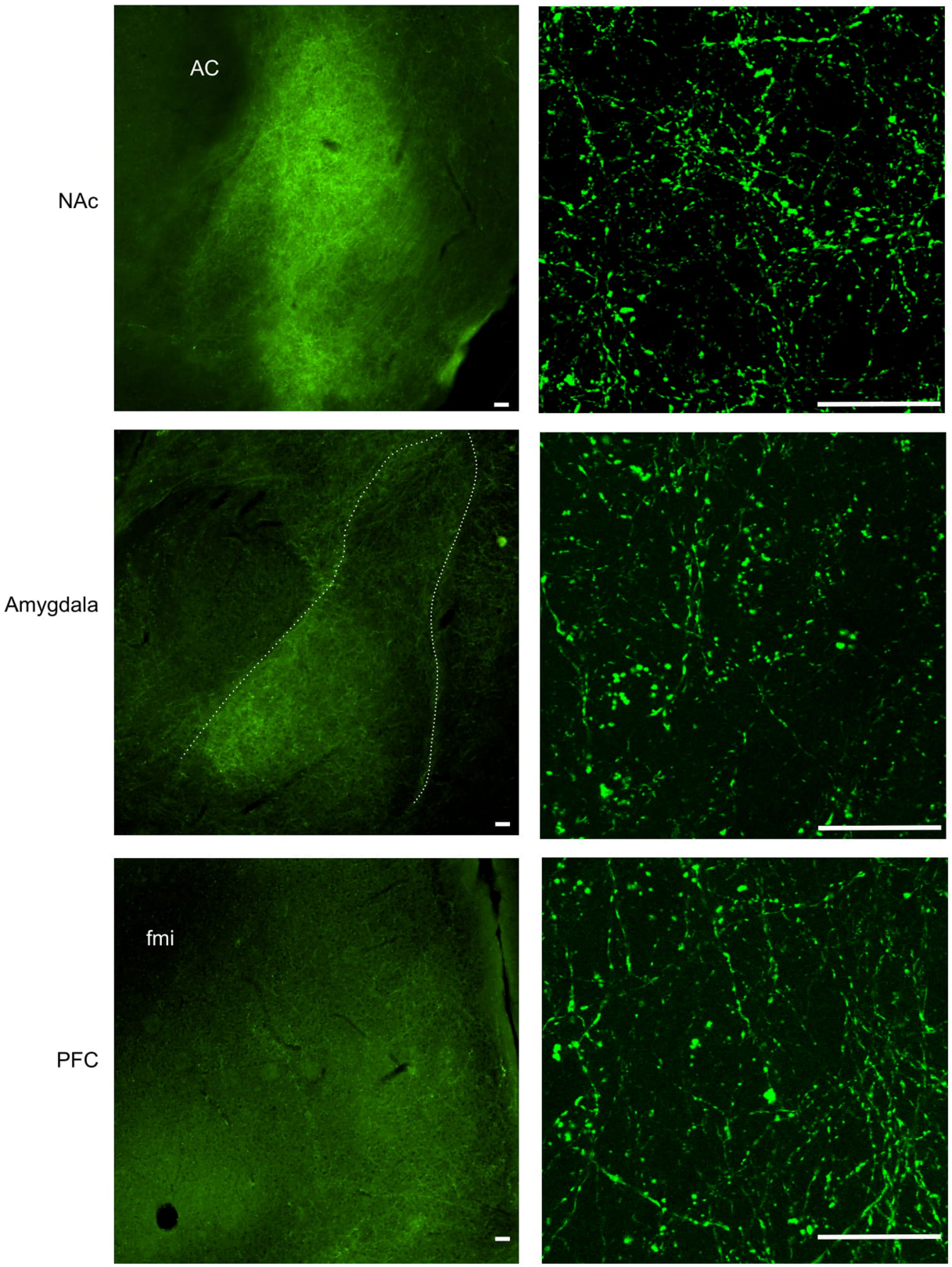
Extended Data Fig. 4 | D1 receptor signalling is required for LTP induction at non-specific NAc synapses. a, Pre-incubation with D1 receptor antagonist SCH23390 blocks LTP induction in response to HFS. **b**, Summary data from the last 5 min of recording. Two-tailed Mann–Whitney *U*-test: $U = 2$, $P = 0.0317$, $n = 5$ mice per group. $^{\#}$Two-tailed paired *t*-test baseline EPSC amplitude/30 min post HFS: control: $t = 3.017$, $P = 0.0393$, $n = 5$ cells; SCH: $t = 0.5016$, $P = 0.6372$, $n = 6$ cells. LTP kinetic data are plotted in 1-min bins. Centre values represent mean and error bars represent s.e.m. For box plots, the line in the middle of the box is plotted at the median. The box extends from the 25th to 75th percentiles. Whiskers represent minimum and maximum.

© 2018 Springer Nature Limited. All rights reserved.



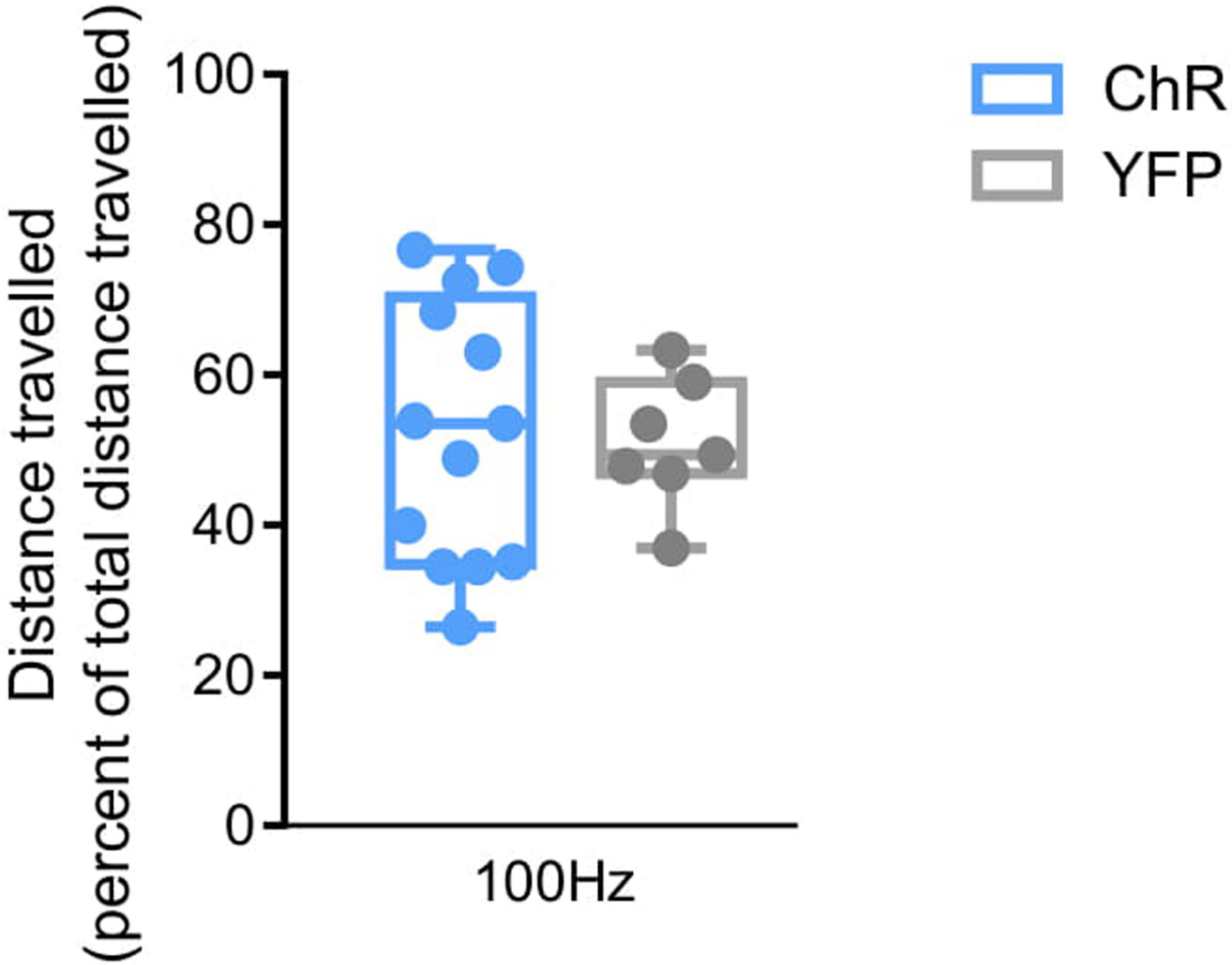
Extended Data Fig. 5 | D2 receptors are not required for LTP induction. **a**, Pre-incubation with the D2-receptor antagonist sulpiride does not affect the ability to elicit LTP in response to HFS in D2R-MSNs. **b**, Summary data from the last 5 min of recording. Two-tailed Mann–Whitney *U*-test: $U = 20$, $P = 0.9452$, $n = 6$, 7 mice. #Two-tailed paired *t*-test baseline EPSC amplitude/30 min post HFS: control: $t = 3.840$, $P = 0.0121$, $n = 6$ cells; sulpiride: $t = 4.246$, $P = 0.0022$, $n = 10$ cells. **c**, Representative traces of EPSCs from control and sulpiride-treated D2R-MSNs. #Significant increase in EPSC amplitude above baseline revealed by paired *t*-test. LTP kinetic data are plotted in 1-min bins. Centre values represent mean and error bars represent s.e.m. For box plots, the line in the middle of the box is plotted at the median. The box extends from the 25th to 75th percentiles. Whiskers represent minimum and maximum. Scale bar for representative traces, 10 pA/10 ms.

© 2018 Springer Nature Limited. All rights reserved.



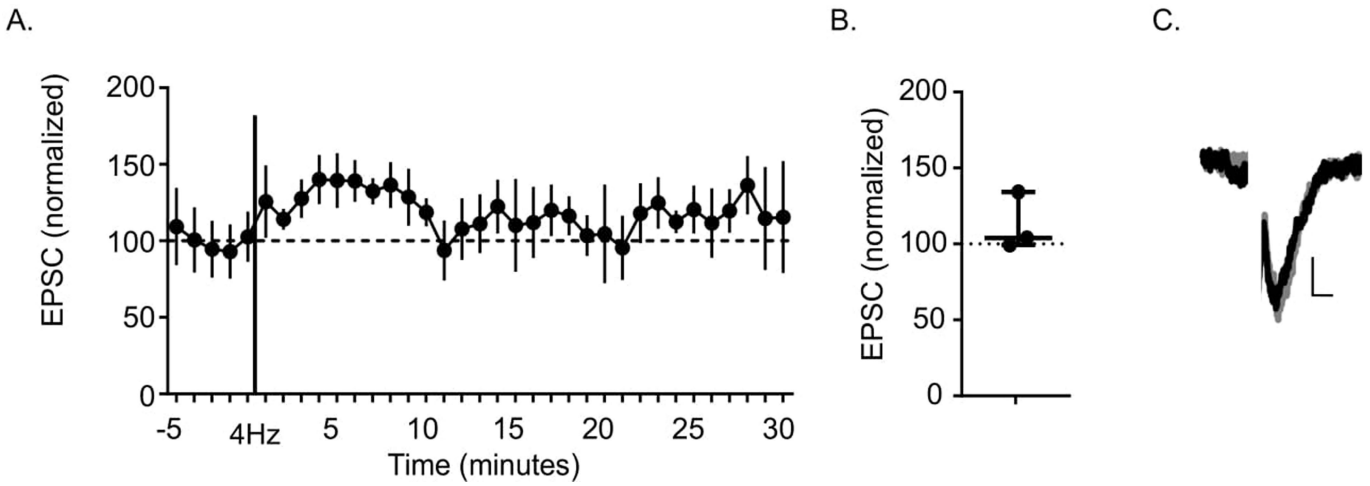
Extended Data Fig. 6 | Collaterals from hippocampus–NAc projecting cells. Representative images of labelled hippocampal fibres. Hippocampal cells projecting to the NAc were labelled by injecting a retrograde virus expressing Cre recombinase into the shell of the NAc and a Cre-dependent virus containing YFP into the ventral hippocampus. Some collaterals are visible in the amygdala as well as the prelimibic and infralimbic regions of the PFC. Right: 100× image showing labelling of fibres. Scale bars, 50 μm. AC, anterior commissure; fmi, forceps minor of the corpus callosum. This was replicated in one other mouse.

© 2018 Springer Nature Limited. All rights reserved.



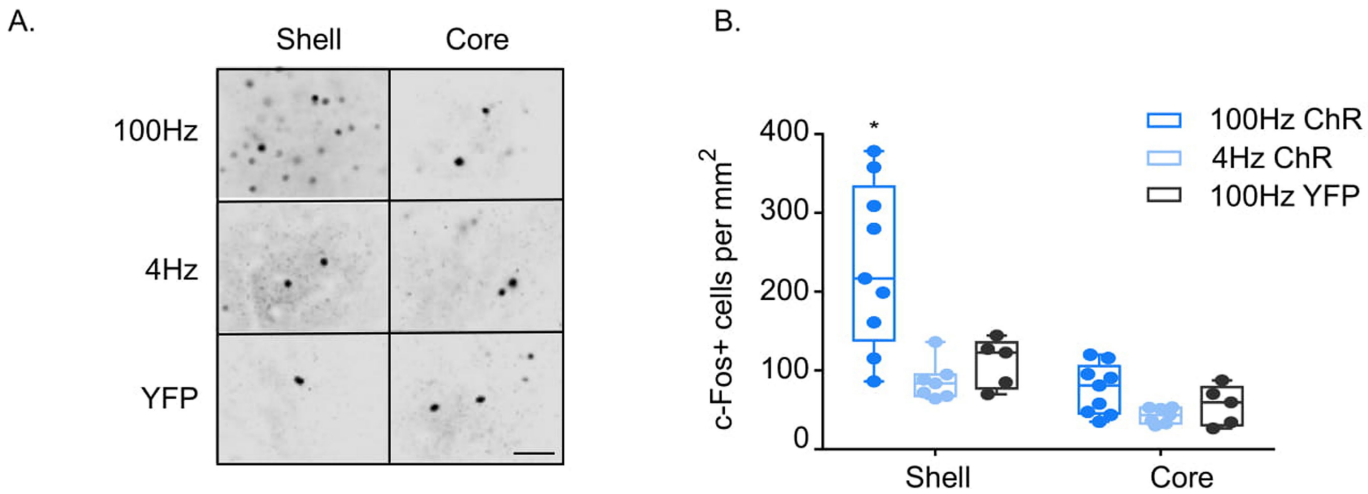
Extended Data Fig. 7 | HFS does not alter locomotor activity. Distance travelled during the conditioning segment of the CPP paradigm. Data were normalized to the distance the mouse travelled during the 'no stimulation' portion of the test. Two-tailed Mann–Whitney *U*-test: $U = 43$, $P = 0.8773$, $n = 13$, 7 mice. The line in the middle of the box is plotted at the median. The box extends from the 25th to 75th percentiles. Whiskers represent minimum and maximum.

© 2018 Springer Nature Limited. All rights reserved.



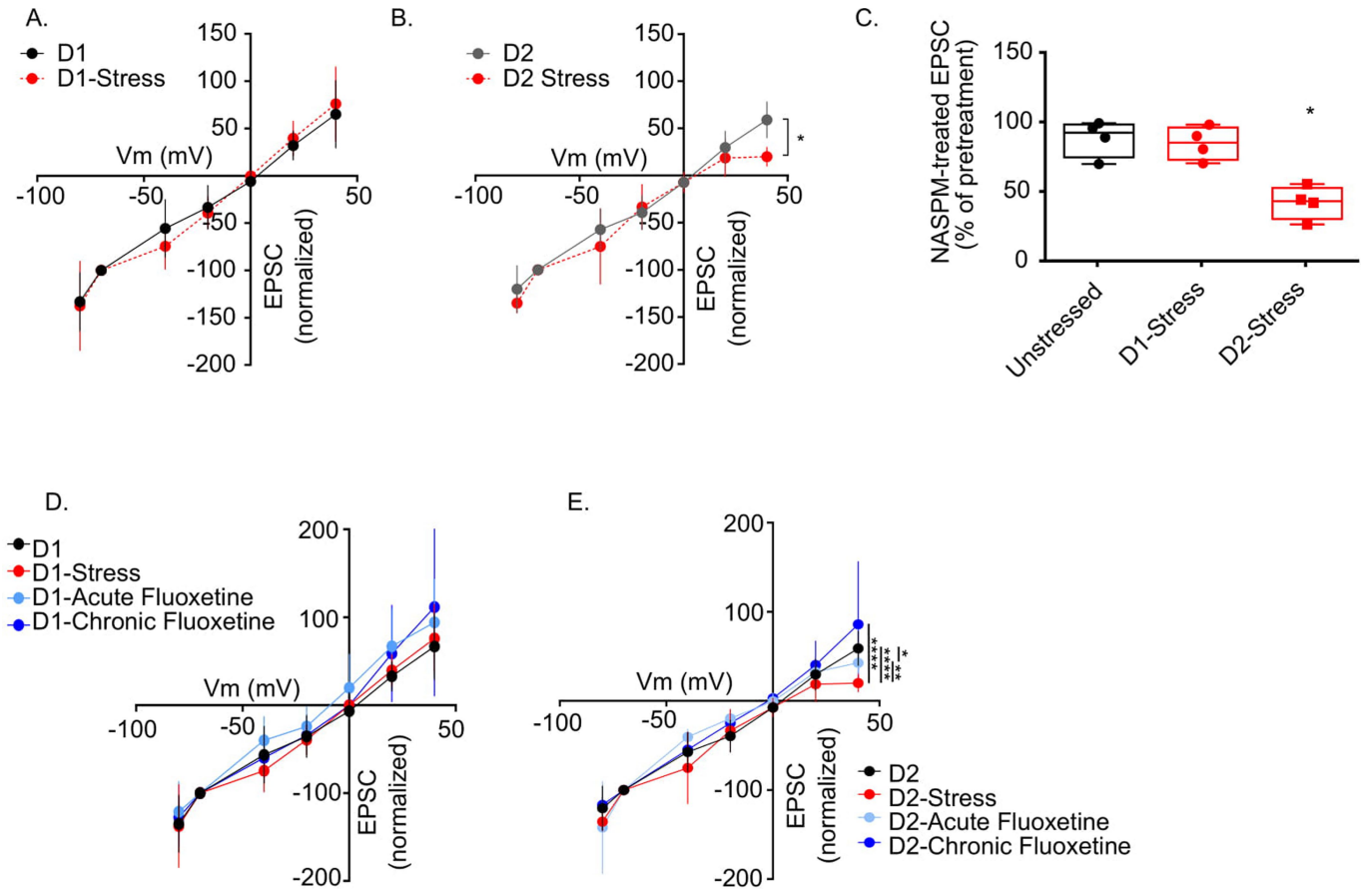
Extended Data Fig. 8 | Stimulation at 4 Hz does not induce LTP. **a**, Stimulation at 4 Hz does not potentiate EPSCs. Data are plotted in 1-min bins. Centre values represent mean and error bars represent s.e.m. **b**, Summary data from the last 5 min of recording. Two-tailed paired *t*-test: $t = 1.171$, d.f. $= 2$, $P = 0.3621$, $n = 3$ cells. The line in the middle of the box is plotted at the median. The box extends from the 25th to 75th percentiles. Whiskers represent minimum and maximum. Scale bar for representative traces, 10 pA/10 ms.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | HFS induces c-Fos expression in the NAc shell. **a**, Representative images from NAc core and shell. Black/grey dots represent c-Fos-positive cells. Scale bar, 50 μm. **b**, Stimulation at 100 Hz but not 4 Hz increases c-Fos expression in the NAc shell. Two-way ANOVA: $F_{2,36} = 5.262$, $P = 0.0099$, $n = 9, 5, 7$ mice.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Chronic stress leads to preferential insertion of GluA2-lacking AMPA receptors in D2R-MSNs. a, Chronic stress does not alter subunit composition in D1R-MSNs. Two-tailed Mann–Whitney *U*-test of amplitude at +40 mV: $U = 35$, $P = 0.6665$, $n = 6$, 7 mice. **b**, D2R-MSNs from mice exposed to chronic stress show inward rectification at positive membrane potentials. Two-tailed Mann–Whitney *U*-test of amplitude at +40 mV: $U = 0$, $P = 0.0006$, $n = 6$, 7 mice. **c**, NASPM decreases EPSC amplitude in D2R-MSNs from mice exposed to chronic stress. Kruskal–Wallis test: $H = 7.423$, $P = 0.0132$, $n = 4$ mice per group. **d**, Current–voltage relationships in D1R-MSNs remain unaffected by chronic stress or fluoxetine treatment. Kruskal–Wallis test with Dunn's post hoc test: $H = 0.9436$, $P = 0.8149$, $n = 5, 5, 5, 4$ mice. **e**, D2R-MSNs from mice exposed to chronic stress treated with chronic fluoxetine show a linear current–voltage relationship, similar to unstressed controls. Inward rectification is observed in D2R-MSNs from mice exposed to chronic stress alone or chronic stress with acute fluoxetine treatment. Kruskal–Wallis test with Dunn's post hoc test: $H = 31.42$, $P < 0.0001$, $n = 5, 5, 5, 8$ mice. The line in the middle of the box is plotted at the median. The box extends from the 25th to 75th percentiles. Whiskers represent minimum and maximum. Centre values represent mean and error bars represent s.e.m. $^{****}P < 0.0001$, $^{**}P < 0.01$, $^{*}P < 0.05$.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0753-3

Trophoblast organoids as a model for maternal–fetal interactions during human placentation

Margherita Y. Turco[1,2,3]*, Lucy Gardner[1,3], Richard G. Kay[4], Russell S. Hamilton[2,3], Malwina Prater[2,3], Michael S. Hollinshead[1], Alasdair McWhinnie[5], Laura Esposito[1], Ridma Fernando[2,3], Helen Skelton[1], Frank Reimann[4], Fiona M. Gribble[4], Andrew Sharkey[1,3], Steven G. E. Marsh[5,6], Stephen O'Rahilly[4], Myriam Hemberger[3,7,8], Graham J. Burton[2,3,9]* & Ashley Moffett[1,3,9]*

The placenta is the extraembryonic organ that supports the fetus during intrauterine life. Although placental dysfunction results in major disorders of pregnancy with immediate and lifelong consequences for the mother and child, our knowledge of the human placenta is limited owing to a lack of functional experimental models[1]. After implantation, the trophectoderm of the blastocyst rapidly proliferates and generates the trophoblast, the unique cell type of the placenta. In vivo, proliferative villous cytotrophoblast cells differentiate into two main sub-populations: syncytiotrophoblast, the multinucleated epithelium of the villi responsible for nutrient exchange and hormone production, and extravillous trophoblast cells, which anchor the placenta to the maternal decidua and transform the maternal spiral arteries[2]. Here we describe the generation of long-term, genetically stable organoid cultures of trophoblast that can differentiate into both syncytiotrophoblast and extravillous trophoblast. We used human leukocyte antigen (HLA) typing to confirm that the organoids were derived from the fetus, and verified their identities against four trophoblast-specific criteria[3]. The cultures organize into villous-like structures, and we detected the secretion of placental-specific peptides and hormones, including human chorionic gonadotropin (hCG), growth differentiation factor 15 (GDF15) and pregnancy-specific glycoprotein (PSG) by mass spectrometry. The organoids also differentiate into HLA-G$^+$ extravillous trophoblast cells, which vigorously invade in three-dimensional cultures. Analysis of the methylome reveals that the organoids closely resemble normal first trimester placentas. This organoid model will be transformative for studying human placental development and for investigating trophoblast interactions with the local and systemic maternal environment.

To devise an organoid culture system suitable for trophoblast, we focused on maternal and placental products that might signal to the stem and progenitor cells that reside in areas of Ki67$^+$ villous cytotrophoblast (VCT) proliferation and/or at the base of the cytotrophoblast cell columns (CCCs) that give rise to extravillous trophoblast cells (EVT)[4–9] (Fig. 1a, b). We investigated signalling pathways between 6 and 8 weeks gestation when proliferation is high: WNT through β-catenin; TGFβ through SMAD2 and SMAD3; and MAPK through ERK1, ERK2 and STAT3. Our findings led to empirical trials of agonists and antagonists, along with other agents, which resulted in a basal trophoblast organoid medium (TOM) composed of EGF, FGF2, CHIR99021 (a WNT activator), A83-01 (a TGFβ and SMAD inhibitor) and R-spondin 1 (Extended Data Fig. 1a–d). To prepare isolates of trophoblast cells, first trimester placentas (6–9 weeks gestation) are enzymatically digested to enrich for cell clusters that express a marker of proliferative trophoblast, EPCAM[10] (Fig. 1c). Cell clusters are seeded into Matrigel drops and grown in TOM. Although some growth is seen, we also tested factors used in other organoid systems and/or present in the first trimester microenvironment (Extended Data

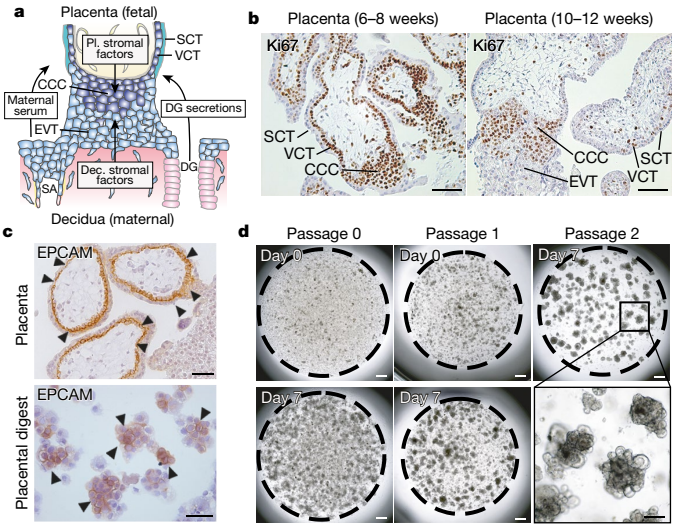


Fig. 1 | Establishment of long-term organoid cultures of trophoblast from human placentas. a, Schematic of a placental villus at the maternal–fetal interface in the first trimester of pregnancy showing the different trophoblast subsets: SCT, VCT, CCC and EVT. Sources of the intrinsic and extrinsic signals that could signal to proliferative Ki67$^+$ trophoblast cells (dark blue) are shown. dec, decidual; DG, decidual gland; SA, spiral artery; Pl, placental. **b**, Immunohistochemical staining for Ki67 in early first trimester placenta (6–8 weeks gestation) compared to late first trimester placenta (10–12 weeks gestation). The proportion of proliferative cells is greatly reduced towards the end of the first trimester, and the cells are localized mostly in the CCCs. Representative images from $n = 6$ for each tissue type. Scale bars, 100 μm. **c**, Immunohistochemical staining for EPCAM in first trimester placenta (6–8 weeks gestation) and cell clusters from first trimester placental digests. The experiment was independently repeated twice with similar results. Arrowheads show that the VCT and CCCs are EPCAM$^+$, and these cells are present in the cell clumps from the placental digests. Scale bars, 50 μm (placenta) and 200 μm (placental digest). **d**, Time course for derivation of trophoblast organoids from one placental isolate. Bright-field images of Matrigel drops after seeding placental digests starting from passage 0 (day 0), until the generation of homogenous trophoblast organoids (passage 2, day 7). For passages 0 and 1, time points at days 0 and 7 are shown. Passage 2 (day 7) is shown together with a magnified view of trophoblast organoids (boxed area). The experiment was independently repeated for all organoid cultures with similar results. Scale bars, 500 μm and 200 μm (magnified view).

[1]Department of Pathology, University of Cambridge, Cambridge, UK. [2]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. [3]Centre for Trophoblast Research, University of Cambridge, Cambridge, UK. [4]Metabolic Research Laboratories and MRC Metabolic Diseases Unit, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. [5]Anthony Nolan Research Institute, Royal Free Hospital, London, UK. [6]UCL Cancer Institute, Royal Free Campus, London, UK. [7]Epigenetics Programme, The Babraham Institute, Cambridge, UK. [8]Present address: Department of Biochemistry and Molecular Biology, University of Calgary, Calgary, Alberta, Canada. [9]These authors jointly supervised this work: Graham J. Burton, Ashley Moffett. *e-mail: myt25@cam.ac.uk; gjb2@cam.ac.uk; am485@cam.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

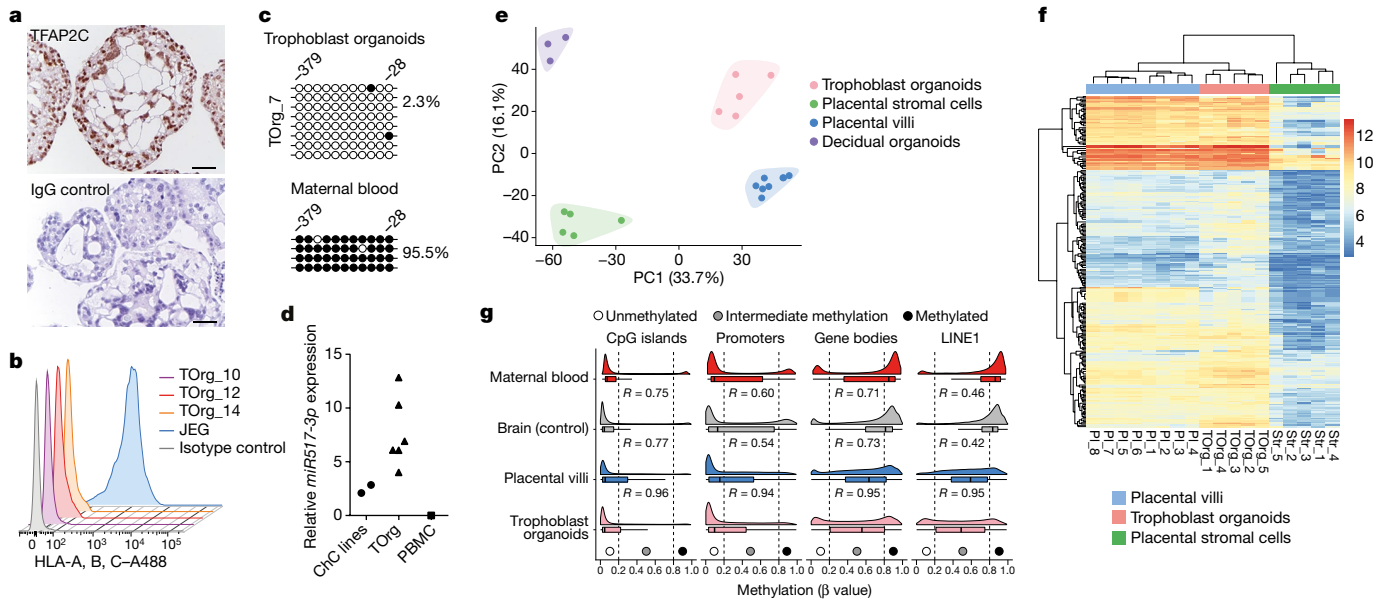


Fig. 2 | Trophoblast organoids retain characteristic features of first trimester trophoblast in vivo, and similar transcriptomic and global methylation profiles. a, Immunohistochemical staining for TFAP2C shows uniform expression (representative images from $n = 20$). Scale bars, 50 μm. **b**, FACS analysis of three trophoblast organoids (TOrg_10, TOrg_12 and TOrg_14) and JEG-3 cells (positive control) with the Alexa-488-conjugated monoclonal antibody clone W6/32, which binds all HLA class I molecules. See Extended Data Fig. 3c for the gating strategy. The experiment was independently repeated three times. **c**, Bisulfite sequencing of the *ELF5* promoter region of trophoblast organoids (TOrg_7) and matched maternal leukocytes (positive control). The relative percentage of methylated cytosine residues (filled circles) is indicated. **d**, qPCR analysis for *miR517-3p* from the C19MC cluster on trophoblast organoids ($n = 6$), JEG-3 and JAR choriocarcinoma (ChC) cell lines (positive controls) and peripheral blood monocytes (PBMC) (low expression/negative control). The graph shows relative expression levels to the housekeeping gene *RNU48*. **e**, Principal component (PC) analysis of placental villi ($n = 8$); trophoblast organoids derived from different placentas, TOrg_1 (passage 4), TOrg_2 (passage 7), TOrg_3 (passage 6), TOrg_4 (passage 4) and TOrg_5 (passage 6) ($n = 5$); placental stromal cells ($n = 5$); and decidual organoids ($n = 3$). Analysis based on 12,673 probes. Organoids cluster more closely with placental villi on the PC1 axis. **f**, Clustered heat map of differentially expressed genes in first trimester placental villi ($n = 8$) (blue), trophoblast organoids ($n = 5$) (pink) and cultured placental villous stromal cells ($n = 5$) (green). **g**, Distribution of methylcytosine across genomic features is similar between trophoblast organoids ($n = 4$) and placental samples ($n = 4$). By contrast, the brain ($n = 1$) and maternal blood ($n = 5$) samples show distinct patterns, especially across CpG islands, gene bodies and LINE1 elements. Pearson's correlation coefficient (R) is indicated for each comparison compared to trophoblast organoid samples (all $P < 2.2 \times 10^{-16}$). Density plots are scaled to area.

Fig. 2a, Supplementary Table 1a). The growth factors HGF, PGE2 and Y-27632 (a ROCK inhibitor) increase cell viability and growth and, when they are combined with TOM, there is rapid expansion of cells within a week (Fig. 1d, Supplementary Table 1b). After the first passage, organoid structures appear and homogeneous trophoblast organoids are established within two passages (10–14 days), with an efficiency of 91% (20 out of 22 patient samples). To confirm their fetal origin, we used microsatellite analysis and HLA typing (Extended Data Table 1). Derivation of the organoids in TOM in the absence of individual media components reveals that EGF is the most important, with effects also seen with Y-27632, A83-01 and CHIR99021. We have now derived trophoblast organoids that are genetically stable after many passages (Extended Data Fig. 2b, c); three randomly selected cultures are still growing after a year (Supplementary Table 1c), are healthy and show active mitochondrial function (Extended Data Fig. 2d). Maternal epithelial cells are always detectable by flow cytometry in placental cell isolates. Nicotinamide enriches for cystic structures that resemble decidual glandular organoids at early stages of derivation (Extended Data Fig. 2a), and microsatellite analysis and HLA typing confirmed their maternal origin (Extended Data Table 2). By selecting the appropriate media, we can derive both decidual glandular and trophoblast organoid cultures from the same pregnancy[11] (Extended Data Fig. 2e). This highlights the importance of verifying the maternal or fetal origin of any cultures derived from decidual or placental cell isolates.

The trophoblast identity of the organoids was verified on the basis of our previously defined criteria: they express the markers GATA3, KRT7, EGFR, TFAP2A and TFAP2C; they lack expression of HLA class I molecules; they express *ELF5* and its promoter is hypomethylated; and they express microRNAs (miRNAs) from the chromosome 19 miRNA cluster (C19MC) at similar or higher levels than the choriocarcinoma lines JEG-3 and JAR[3] (Fig. 2a–d, Extended Data Fig. 3a–f). To assess how well the trophoblast organoids recapitulate their tissue of origin in an unbiased approach, we performed a microarray analysis of established organoids and compared them to first trimester placental villi (also containing stromal, Hofbauer and endothelial cells) and cultured villous stromal cells. To check for maternal cell contamination, decidual glandular organoids were included. The results were analysed by principal component analysis and hierarchical clustering (Fig. 2e, f, Extended Data Fig. 4a). PC1 shows trophoblast organoids cluster closely to the placenta with enrichment for trophoblast-specific genes such as *CGB3*, *GATA3* and *PSG6* compared to the stromal cells and glandular organoids. PC2 highlights epithelial genes (for example *CLDN3*, *TACSTD2* and *KRT23*) that are specific for trophoblast compared to other placental cells (Extended Data Fig. 4b). Immunohistochemical analysis of placental villi and trophoblast organoids confirms that KRT23 is a trophoblast-specific keratin (Extended Data Fig. 4c). Comparison of differentially expressed genes (fold change ≥ 2, adjusted $P \leq 0.05$) between placental villi, trophoblast organoids and stromal cells highlights other genes of interest, such as *PGF*, *CCNE1*, *ERBB3* and *FOLR1*; the translation of CCNE1 in trophoblast was validated by immunohistochemistry (Extended Data Fig. 4d, e). The imprinted genes *PEG3* and *PEG10* are also highly expressed in the trophoblast organoids (Extended Data Fig. 4d). Besides the known transcription factor genes *GATA3* and *TFAP2C*, differentially expressed genes emerged such as *ELF3* (Extended Data Fig. 5a, b). Genome-wide methylation analysis revealed a high degree of correlation between trophoblast organoids and first trimester placental villi across different genomic elements compared to blood and brain (Fig. 2g). The hypomethylation of the

© 2018 Springer Nature Limited. All rights reserved.

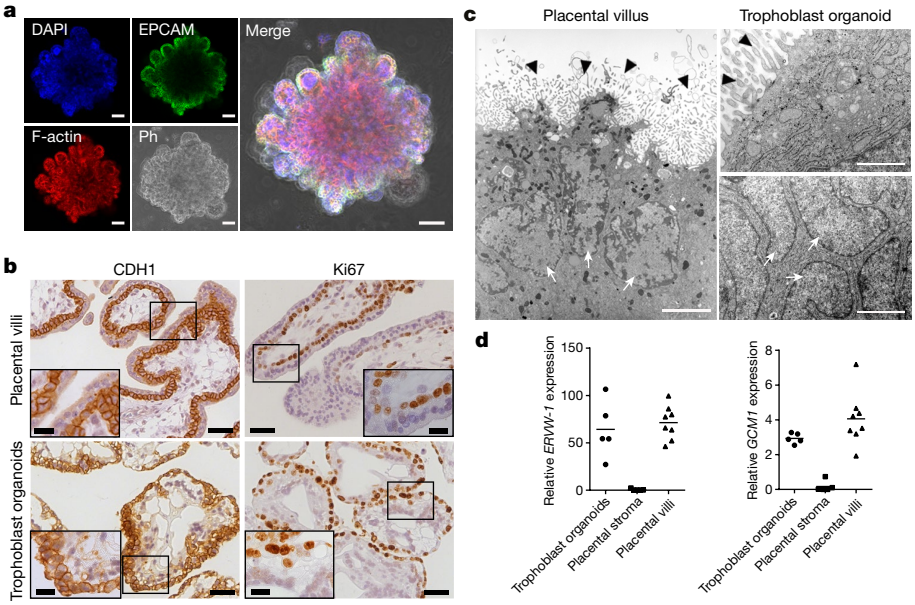


Fig. 3 | Trophoblast organoids form complex structures resembling placental villi with formation of the SCT. a, Confocal microscopy images of trophoblast organoid stained for F-actin, EPCAM and DAPI, with phase-contrast (Ph) and merged images (representative image from $n = 5$). The EPCAM$^+$ cells are the outer cells of the organoid. Scale bars, 100 μm. **b**, Immunohistochemistry staining for CDH1 and Ki67 of first trimester placenta and trophoblast organoids (representative images from $n = 6$ (placental villi) and $n = 20$ (trophoblast organoids). VCT cells stain positively for CDH1. Ki67 is present in the inner VCT layer in placental villi and the outer layer in organoids. Scale bars, 50 μm and 20 μm (insets). **c**, Electron micrograph images of first trimester SCT compared to the centre of a trophoblast organoid. Surface microvilli (arrowheads) and multinucleated areas (white arrows) can be seen. Scale bars, 5 μm (placenta), 1 μm (trophoblast organoids, microvilli) and 2.5 μm (trophoblast organoids, nuclei). Representative images from $n = 2$. **d**, qPCR analysis of *ERVW-1* and *GCM1* genes in trophoblast organoids ($n = 5$) compared to whole placental villi ($n = 8$) and placental stromal cells ($n = 5$). Graphs show expression levels relative to geometric mean of the housekeeping genes *TBP*, *TOP1* and *HPRT1*. Horizontal lines denote the mean expression levels.

ELF5 promoter was also confirmed (Extended Data Fig. 5c, d). Analysis of the promoter regions of genes with similar methylation patterns to *Elf5* in mouse trophoblast stem cells shows that *ELF5*, *EZR*, *TINAGL1* and *LASP1* are similarly hypomethylated in human placental villi and trophoblast organoids[12] (Extended Data Fig. 5e, f). Gene Ontology analysis of differentially expressed genes, represented by a chord diagram, shows terms that describe metabolic processes and cell–cell organization that converge on epithelial, developmental and hormonal pathways, and identified the genes *FZD5* (WNT signalling), *INSIG1* (insulin signalling), *DHCR7* (cholesterol synthesis) and *OCLN* (polarity) (Extended Data Fig. 5g).

Trophoblast organoids grow as complex structures that closely recapitulate the organisation of placental villi in vivo, where VCT stain for EPCAM and CDH1 (Fig. 3a, b). The basement membrane is on the outside in contact with Matrigel, and syncytial masses line the central cavity (Extended Data Fig. 6a). Similar to in vivo, VCT cells are Ki67$^+$ and TP63$^+$ (Fig. 3b, Extended Data Fig. 6b, c). After incubation with EdU, approximately 30–40% of cells are proliferating when the organoids are small (100–200 μm diameter), with a notable decrease as they enlarge and differentiate (Extended Data Fig. 6d). Expression of the syncytiotrophoblast (SCT) markers CD46 and CD71 is detected inside the organoids (Extended Data Fig. 6e). The characteristic features of SCT were confirmed by electron microscopy, in which multinucleated areas with abundant secretory organelles and surface microvilli were detected (Fig. 3c). Lacunae present within the syncytial areas resemble those found in vivo (Extended Data Fig. 6f, g). *GCM1* drives the fusion of VCT into SCT by upregulating *ERVW-1* (also known as *SYNCYTIN-1*)[13,14]. Using quantitative PCR (qPCR), we detected high expression levels of *GCM1* and *ERVW-1* in trophoblast organoids, comparable to those in the placental villi (Fig. 3d). Thus, trophoblast organoids closely mimic the villous placenta both structurally and phenotypically. The SCT secretes proteins and hormones into the maternal systemic circulation that mediate maternal adaptations to pregnancy. We explored the secretory activity of trophoblast organoids using an unbiased proteomic analysis of the conditioned organoid medium by liquid chromatography–tandem mass spectrometry (LC–MS/MS) technology (Fig. 4a). Among the most abundant peptides are placental-specific PSGs and INSL4, the functions of which

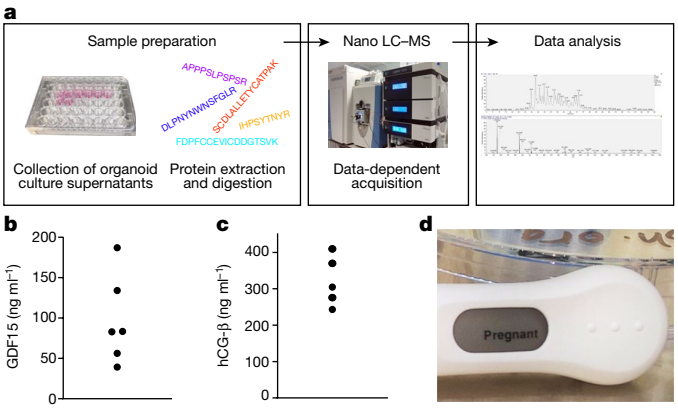


Fig. 4 | The secretome of trophoblast organoids contains placental hormones and proteins. a, Experimental workflow for LC–MS/MS analysis of the secretome of trophoblast organoids. Supernatants from six trophoblast organoid cultures derived independently from six different placental samples were analysed: TOrg_2 (passage 23), TOrg_3 (passage 20), TOrg_5 (passage 6), TOrg_10 (passage 12), TOrg_12 (passage 4) and TOrg_14 (passage 5). **b**, ELISA for GDF15 secreted by trophoblast organoids ($n = 6$). The amount of GDF15 (ng ml^{-1}) produced by trophoblast organoids (between days 7 and 10 after passaging) in 48 h is shown. **c**, ELISA for hCG-β secreted by trophoblast organoids ($n = 5$). The amount of hCG-β (ng ml^{-1}) produced by trophoblast organoids (between days 7 and 10 after passaging) in 48 h is shown. **d**, Over-the-counter pregnancy test denoting 'pregnant' after being placed into a dish containing cultures of trophoblast organoids. Image reproduced with the permission of SPD Swiss Precision Diagnostics GmbH (SPD). Experiment was independently repeated twice.

© 2018 Springer Nature Limited. All rights reserved.

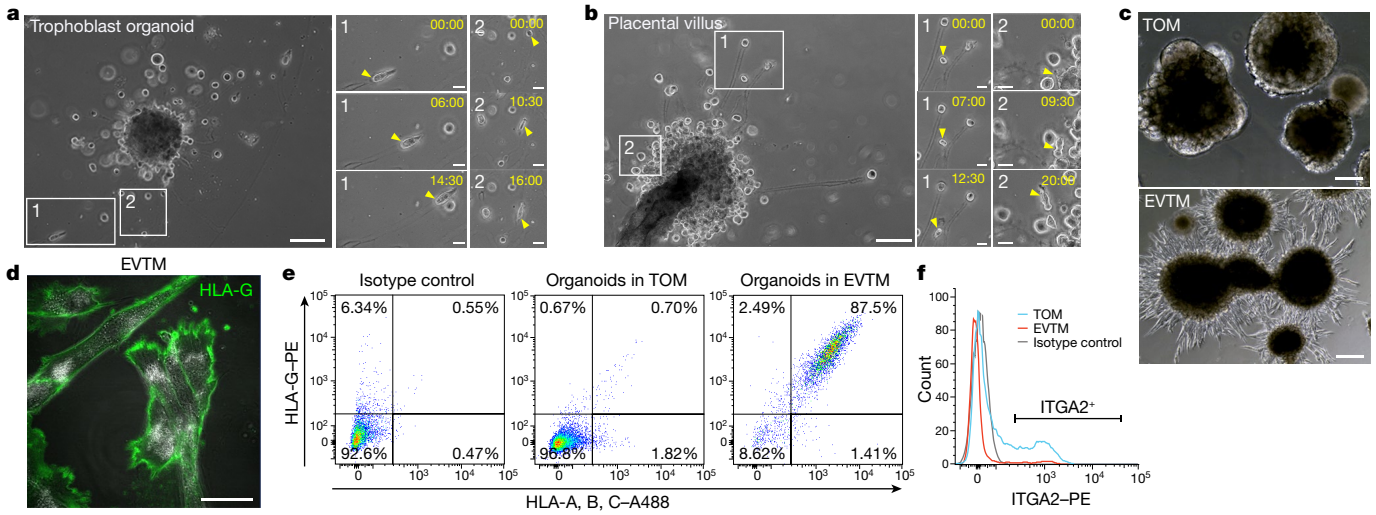


Fig. 5 | Generation of migratory and invasive HLA-G$^+$ EVT cells from trophoblast organoids. a, **b**, Phase-contrast images taken across several *z*-stacks combined into a single image by the extended focus module from Zeiss Axiovision of a trophoblast organoid (**a**) and a placental villous explant (**b**) plated into Matrigel after 7–10 days in EVT differentiation medium (EVTM). Regions of interest are boxed in white and corresponding higher power snapshots are shown with relative time-lapse intervals in yellow (h:min). Migratory cells are labelled by yellow arrowheads. In EVTM, cells from both the organoids and primary tissue show random migration. See also control image for organoid in TOM (Extended Data Fig. 7b) and time-lapse videos (Supplementary Videos 1–6). Scale bars, 200 μm and 50 μm (insets). **c**, Phase-contrast images of trophoblast organoids plated in Matrigel drop and exposed to either TOM (top) or EVTM (bottom). Cells stream out of organoids, digesting the Matrigel and eventually adhering to the plastic only when cultured in EVTM. Cultures after 14–21 days in EVTM. Scale bars, 200 μm. **d**, Live cells growing out from trophoblast organoids in EVTM stained with the HLA-G monoclonal antibody G233. Scale bar, 50 μm. **e**, Flow cytometry of trophoblast organoids cultured in TOM or EVTM and double-stained with monoclonal antibodies W6/32 (all HLA class I molecules) and MEMG9 (specific for HLA-G). In TOM, almost all cells lack HLA class I expression but become HLA-G$^+$ after culturing in EVTM. **f**, Histogram showing organoids cultured in TOM or EVTM stained for ITGA2, which marks cells at the base of the cytotrophoblast cell columns. Before exposure to EVTM, 23% of cells are ITGA2$^+$ but very few are present after differentiation to HLA-G$^+$ EVT. Experiments with placental villous explants were repeated independently twice. All other experiments have been repeated independently at least three times.

are unknown (Extended Data Table 3, Supplementary Table 2a–h). Aldose reductase, which converts glucose to sorbitol, is also detected. High concentrations of sorbitol are present in first trimester placentas[15]. Hence, the organoids also mimic the villous placenta metabolically as well as endocrinologically. Peptides that induce physiological and metabolic adaptations during pregnancy, including hCG, KISS1 and CSH1, are all abundant, in addition to GDF15, which is implicated in hyperemesis gravidarum[16]. GDF15 and hCG are detected by ELISA, showing that full-length and appropriately folded hormones are secreted by trophoblast organoids (Fig. 4b, c). Indeed, the 'pregnant' secretome of the organoids is evident using an over-the-counter pregnancy test kit (Fig. 4d).

Human trophoblast also differentiates to EVT, a process that is crucial for proper placentation. EVT cells express HLA-G and invade decidual tissue to transform the spiral arteries[17]. In TOM, our trophoblast organoids show only sporadic HLA-G$^+$ cells (Extended Data Fig. 7a). Long-term, two-dimensional monolayer cultures of human trophoblast cells derived from first trimester placentas that can differentiate into SCT and EVT were recently described[18]. By adapting this EVT differentiation protocol and culturing both our trophoblast organoids and primary villous explants in their EVT differentiation medium (EVTM)[18], HLA-G$^+$ cells that migrate out of the organoids emerge, digest the Matrigel to form tracks, and eventually adhere to the plastic (Fig. 5a–e, Extended Data Fig. 7b, Supplementary Videos 1–6). In vivo, EVT are generated at the base of the CCCs, where cells express ITGA2[10]. We used flow cytometry to confirm that, after exposure of organoids to EVTM, HLA-G$^+$ EVT appear and ITGA2$^+$ cells disappear (Fig. 5e, f).

In summary, we describe the generation of human trophoblast organoids that grow as complex three-dimensional structures with the fusion of VCT to hCG-secreting SCT, and anatomically and functionally closely resemble the villous placenta in vivo. In addition, we show differentiation to HLA-G$^+$ EVT cells that vigorously invade and digest Matrigel in 3D cultures. After the submission of our paper, a report was published that describes the generation of trophoblast organoids from pooled patient samples; however, to our knowledge, these cannot be cultured long-term and have not been fully characterized using the four trophoblast criteria[19]. Our results complement previous findings and mean that there are now two culture systems (2D and 3D) for human trophoblast[18]. Our 3D model has the advantage of organizing into villous structures that will allow analysis of morphogenetic events. We anticipate that these two different models will provide valuable tools across a range of disciplines. They can be used to study maternal–fetal transmission of xenobiotics, drugs and pathogens, and the proteins and hormones derived from the SCT[20]. Analysis of CCC formation and EVT differentiation in vitro will allow investigation of the decidual microenvironment on trophoblast function, such as the effect of glandular histotrophic nutrition and the influence of the distinctive uterine natural killer cells[21,22]. Major unexplained disorders of pregnancy such as pre-eclampsia, still-birth and fetal growth restriction have their origins in aberrant placental development in the first trimester[23]. Trophoblast organoids can be used to study maternal–fetal interactions after implantation, and the maternal physiological, metabolic and hormonal changes that occur during pregnancy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0753-3.

Received: 3 April 2018; Accepted: 31 October 2018;
Published online 28 November 2018.

1. Burton, G. J. & Jauniaux, E. What is the placenta? *Am. J. Obs. Gynecol* **213**, S6.e1–S6.e8 (2015).
2. Hamilton, W. J. & Boyd, J. D. Development of the human placenta in the first three months of gestation. *J. Anat.* **94**, 297–328 (1960).
3. Lee, C. Q. et al. What is trophoblast? A combination of criteria define human first-trimester trophoblast. *Stem Cell Reports* **6**, 257–272 (2016).

© 2018 Springer Nature Limited. All rights reserved.

4. Kunath, T. et al. Developmental differences in the expression of FGF receptors between human and mouse embryos. *Placenta* **35**, 1079–1088 (2014).
5. Paiva, P. et al. Human chorionic gonadotrophin regulates FGF2 and other cytokines produced by human endometrial epithelial cells, providing a mechanism for enhancing endometrial receptivity. *Hum. Reprod.* **26**, 1153–1162 (2011).
6. Hempstock, J., Cindrova-Davies, T., Jauniaux, E. & Burton, G. J. Endometrial glands as a source of nutrients, growth factors and cytokines during the first trimester of human pregnancy: a morphological and immunohistochemical study. *Reprod. Biol. Endocrinol.* **2**, 50 (2004).
7. Mühlhauser, J. et al. Differentiation and proliferation patterns in human trophoblast revealed by c-erbB-2 oncogene product and EGF-R. *J. Histochem. Cytochem.* **41**, 165–173 (1993).
8. Sonderegger, S., Husslein, H., Leisser, C. & Knofler, M. Complex expression pattern of Wnt ligands and frizzled receptors in human placenta and its trophoblast subtypes. *Placenta* **28** (Suppl.), S97–S102 (2007).
9. Haider, S., Kunihs, V., Fiala, C., Pollheimer, J. & Knöfler, M. Expression pattern and phosphorylation status of Smad2/3 in different subtypes of human first trimester trophoblast. *Placenta* **57**, 17–25 (2017).
10. Lee, C. Q. E. et al. A niche of trophoblast progenitor cells identified by integrin α2 is present in first trimester human placentas. *Development* **145**, (2018).
11. Turco, M. Y. et al. Long-term, hormone-responsive organoid cultures of human endometrium in a chemically defined medium. *Nat. Cell Biol.* **19**, 568–577 (2017).
12. Cambuli, F. et al. Epigenetic memory of the first cell fate decision prevents complete ES cell reprogramming into trophoblast. *Nat. Commun.* **5**, 5538 (2014).
13. Yu, C. et al. GCMa regulates the syncytin-mediated trophoblastic fusion. *J. Biol. Chem.* **277**, 50062–50068 (2002).
14. Frendo, J. L. et al. Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation. *Mol. Cell. Biol.* **23**, 3566–3574 (2003).
15. Jauniaux, E., Hempstock, J., Teng, C., Battaglia, F. C. & Burton, G. J. Polyol concentrations in the fluid compartments of the human conceptus during the first trimester of pregnancy: maintenance of redox potential in a low oxygen environment. *J. Clin. Endocrinol. Metab.* **90**, 1171–1175 (2005).
16. Petry, C. J. et al. GDF15 concentrations in maternal serum associated with vomiting in pregnancy: the Cambridge baby growth study. Preprint at https://www.bioRxiv.org/content/early/2017/11/17/221267 (2017).
17. Pijnenborg, R., Vercruysse, L. & Hanssens, M. The uterine spiral arteries in human pregnancy: facts and controversies. *Placenta* **27**, 939–958 (2006).
18. Okae, H. et al. Derivation of human trophoblast stem cells. *Cell Stem Cell* **22**, 50–63.e6 (2018).
19. Haider, S. et al. Self-renewing trophoblast organoids recapitulate the developmental program of the early human placenta. *Stem Cell Reports* **11**, 537–551 (2018).
20. Robbins, J. R. & Bakardjiev, A. I. Pathogens and the placental fortress. *Curr. Opin. Microbiol.* **15**, 36–43 (2012).
21. Burton, G. J., Jauniaux, E. & Charnock-Jones, D. S. The influence of the intrauterine environment on human placental development. *Int. J. Dev. Biol.* **54**, 303–312 (2010).
22. Moffett, A. & Colucci, F. Co-evolution of NK receptors and HLA ligands in humans is driven by reproduction. *Immunol. Rev.* **267**, 283–297 (2015).
23. Smith, G. C. First-trimester determination of complications of late pregnancy. *J. Am. Med. Assoc.* **303**, 561–562 (2010).

Acknowledgements The authors are grateful to patients for donating tissue for research. We thank D. Moore and staff at the Addenbrooke's Hospital, Cambridge; J. Bauer, E. Clemente, E. Farnell, C. Reitter from Cambridge Genomic Services, Department of Pathology; K. Burling and Core Biochemical Assay Laboratory staff at NIHR Cambridge Biomedical Research Centre; I. Simonic; and T. Cindrova-Davies, H. W. Yung, B. D. Simons, B. K. Koo, M. Huch, Y. W. Loke and all members of the Moffett and Burton laboratories. This work was funded by Centre for Trophoblast Research, University of Cambridge and Medical Research Council (MR/L020041/1). A. Moffett holds a Wellcome Trust investigator award (200841/Z/16/Z). M.Y.T. received funding from EU Seventh Framework Programme for research, technological development and demonstration (PIEF-GA-2013-629785) and now holds a Royal Society Dorothy Hodgkin Fellowship. F.R. and F.M.G. hold Wellcome Trust joint investigator awards (106262/Z/14/Z and 106263/Z/14/Z), and funding from MRC-Metabolic Diseases Unit (MRC_MC_UU_12012/3), NIHR-Cambridge BRC and AZ/MedImmune.

Reviewer information *Nature* thanks M. Roberts and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.Y.T. and L.G. designed and carried out all the core experiments and analysed the data; R.G.K. performed LC–MS/MS analysis; R.S.H. and M.P. performed microarray and DNA methylation analysis; A. McWhinnie and S.G.E.M. performed and analysed the STR and HLA typing; M.S.H. performed electron microscopy and confocal analysis; R.F. and H.S. assisted with IHC and ELISA experiments; L.E. performed qPCR experiments; A.S., F.M.G., F.R., M.H. and S.O. provided intellectual input to the study; M.H. performed the ELF5 methylation experiment; G.J.B. and A. Moffett assisted in analysis and interpretation of results and jointly supervised the project. M.Y.T., G.J.B. and A. Moffett wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0753-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0753-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.Y.T., G.J.B. or A. Moffett.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Patient samples. All tissue samples used for this study were obtained with written informed consent from all participants in accordance with the guidelines in The Declaration of Helsinki 2000. Elective terminations of normal pregnancies were performed at Addenbrooke's Hospital (6–12 weeks gestation) under ethical approval from the Cambridge Local Research Ethics Committee (04/Q0108/23). Human peripheral blood was collected from a healthy donor in a BD vacutainer following informed consent and in accordance with the ethical approval of the Human Biology Research Ethics Committee, University of Cambridge (HBREC.2016.03).

Derivation of trophoblast organoids from human placental tissue. To obtain trophoblast-enriched cell suspensions, villi from first trimester placental tissue were sequentially digested with 0.2% trypsin-250 (Pan Biotech P10-025100P), 0.02% EDTA (Sigma E9884) in PBS, then 1.0 mg ml^{-1} collagenase V (Sigma C9263) in Hams F12/10% FBS. Both digests were pooled, washed in Advanced DMEM/F12 medium (Gibco 12634-010) and re-suspended in approximately 10× volume growth-factor-reduced Matrigel (Corning 356231) on ice. Drops (25 μl) were plated per well into a 48-well culture plate (Costar 3548), set at 37 °C for 15 min and overlaid with 250 μl trophoblast organoid medium (TOM; Supplementary Table 1b). Cultures were maintained in 5% CO_2 in a humidified incubator at 37 °C. Medium was replaced every 2–3 days. Small organoid clusters became visible by around day 7 and were passaged when at least 50% had reached a diameter of 200–300 μm (usually between days 7 and 10). Mechanical disruption was achieved with Eppendorf Explorer Plus automatic pipettes on a mix cycle of 99 rounds (4–5 times), maximum speed. Organoids from the same sample were initiated and maintained in the absence of each individual component to test its importance. Frozen stocks of organoids were made in 70% TOM, 20% FBS and 10% DMSO freezing medium and stored in liquid nitrogen. A step-by-step protocol of the derivation and maintenance of human trophoblast organoid cultures can be found at Nature Protocol Exchange[24].

Generation of EVT cells from trophoblast organoids. Trophoblast organoids were passaged and plated into 35-mm dishes or ibidi μ-dishes (Thistle Scientific 81156). Differentiation was achieved through the modification of a protocol described previously[18]. After passaging, organoids were maintained in TOM for 3–4 days and switched to EVT differentiation medium (EVTM: advanced DMEM/F12, 0.1 mM 2-mercaptoethanol (Gibco 31350), 0.5% penicillin–streptomycin, 0.3% BSA (Sigma A8412), 1% ITS-X supplement (Gibco 51500-056), 100 ng ml^{-1} NRG1 (Cell Signaling 5218SC), 7.5 μM A83-01 (Tocris Biotechne 2939) and 4% knockout serum replacement (ThermoFisher 10828010)). When organoids showed outgrowth of cells (typically days 7–10), the medium was changed to EVTM without NRG1 for a further 7–10 days. For comparison, fresh placental villi were embedded in 300 μl Matrigel in ibidi μ-dishes and grown under the same conditions.

Isolation of placental stromal cells. Placental villous stromal cells were isolated by digesting the tissue remaining after the initial trypsin and collagenase digests in 10–15 ml collagenase V in Ham's F12 containing 10% FBS, with gentle shaking at 37 °C for 5–10 min. The cell suspension was filtered through gauze, washed and pelleted. Cells were resuspended in Advanced DMEM/F12 with 10% FBS and additional L-glutamine, non-essential amino acids (Gibco 11140-035) and primocin (InvivoGen ant-pm-1) and seeded into tissue culture flasks. Cells were cultured to 80–90% confluency and passaged once before use.

Isolation of PBMC. PBMC were isolated from blood by Pancoll-based (Pan-Biotech P04-60500) density gradient separation. PBMC viability was >95% by Trypan Blue exclusion. PBMC were resuspended in QIAzol lysis reagent (Qiagen 79306) for total RNA extraction following the supplier's protocol.

Cell lines. Choriocarcinoma cell lines JEG-3 and JAR (used only as control cell lines for experiments in this study) were purchased from the American Type Culture Collection (ATCC) in 2015. Cells were expanded and frozen stocks were immediately made within a few passages of receipt of these cell lines. Early passage numbers were thawed for this study. They are routinely screened for their unique characteristics: HLA class I expression (JAR are negative) and HLA-G (JEG are positive) and have been tissue typed[3]. JEG-3 and JAR were not tested for mycoplasma upon receipt from the ATCC.

Immunohistochemistry. Organoids were formalin-fixed and embedded as previously described[11]. Immunohistochemistry on sections of organoids and first trimester placentas was performed using heat-induced epitope retrieval buffers (A. Menarini) and Vectastain avidin-biotin-HRP reagents (Vector Lab PK-6100) as previously described[11]. Primary antibodies (Supplementary Table 3) were replaced with equivalent concentrations of isotype-matched mouse or rabbit IgG for controls. Images were captured with a Zeiss Axiovert Z1 microscope and Axiovision imaging software SE64 V4.8.

Immunofluorescence and confocal microscopy. Trophoblast organoids were grown in 4–5 20-μl Matrigel drops in 35-mm ibidi μ-dishes and EdU and/or antibody labelling was performed as previously described[11]. EdU incubation was performed for 1 h at 37 °C in TOM containing 10 μM EdU. For primary and secondary antibodies used, see Supplementary Table 1d. Imaging was carried out using the ZEISS LSM 700 Confocal Laser Scanning Microscope and ZEN Microscope Software.

Mitotracker staining. Mitochondrial function was evaluated by Mitotracker Red-CMXRos (Thermofisher M7512). Organoids were released from Matrigel with Cell Recovery Solution (Corning 354253) and incubated in 500 nM of Mitotracker Red in TOM in suspension at 37 °C for 30 min. The organoids were washed in basal medium, resuspended and plated into a thin layer of Matrigel in ibidi μ-dishes for imaging on a ZEISS LSM 700 Confocal Laser Scanning Microscope with ZEN Microscope Software.

Time-lapse microscopy. Trophoblast organoids or placental villous explants embedded in 300 μl of Matrigel in 35-mm ibidi μ-dishes were imaged at 37 °C by phase-contrast microscopy and across several *z*-stacks on a Zeiss Axiovert Z1 microscope with the multidimensional imaging function of the Axio Observer software Axiovision image software V4.8. The images were compiled into a single video by using the extended focus wavelet function.

ELISA. Conditioned media were collected from organoid cultures and centrifuged to remove debris and stored at −80 °C until use. hCG-β ELISA (Abcam ab108638) was performed on 50 μl supernatant with 100 μl sample buffer in duplicate alongside hCG-β standards following the manufacturer's instructions. The concentration of hCG-β in the supernatants was calculated from the line formula of the standard plots in Microsoft Office Excel. Supernatants were also tested with Clear&Simple Digital Pregnancy Test following the manufacturer's instructions. GDF15 was measured by in-house electrochemiluminescence immunoassay on the MesoScale Discovery assay platform (MSD) using BioTechne DuoSet antibodies and standard (BioTechne DY957). See Supplementary Methods for further details.

Flow cytometry. Organoids were removed from Matrigel with Cell Recovery Solution (Corning 354253) and dissociated with 0.2% trypsin 250 (Pan Biotech P10-025100P), 0.02% EDTA (Sigma E9884) in PBS at 37 °C for 5 min. Cells were washed in medium containing FBS and passed through a 40-μm cell strainer (Falcon 2340). Cells were blocked with human IgG (Sigma I4506) in Dulbecco's PBS (ThermoFisher Scientific 14190136) with 1% FBS before labelling with W6/32–Alexa-488 anti-HLA-A, B, C antibody, HLA-G–PE, ITGA2–PE or isotype-matched controls (Supplementary Table 1d). LIVE/DEAD Fixable Far Red Dead Cell Stain (Life Technologies L10119) was used for live/dead discrimination. Data were acquired using Cytek Development DxP8 (488/637/561). Data were analysed in FlowJo (Tree Star) and all compensation was applied digitally after acquisition.

In situ hybridization assays. In situ hybridization for *LGR5* was performed on 4-μm paraffin sections with the RNAscope 2.0 High definition assay (Advanced Cell Diagnostics) following the manufacturer's instructions. In brief, tissue sections were baked at 60 °C for 1 h, dewaxed with xylene, cleared in 100% ethanol and air-dried before the standard protocol: 10 min in pre-treat buffer 1, 15 min in pre-treat buffer 2 and 30 min at 37 °C in pre-treat buffer 3 followed by incubation with the *LGR5* probe (311021), positive-control probe *UBC* (310041) or negative-control probe *dapB* (310043) for 2 h at 40 °C. The signal was visualized with the amplification kit and DAB for 10 min. Sections were dehydrated, mounted in DPX (Sigma 44581) and imaged on a Zeiss Axiovert Z1 microscope with Axiovision imaging software SE64 V4.8.

Electron microscopy. Trophoblast organoids were directly fixed in 35-mm dishes with 0.5% glutaraldehyde, 0.2 M sodium cacodylate buffer (pH 7.2) for 30 min and post-fixed and reduced with osmium tetroxide as previously described[11]. Ultrathin sections were examined in an FEI Tecnai G2 TEM at 80 kV. Images were acquired with MegaView III CCD and Soft Imaging Systems program. Samples from human placentas were fixed by immersion in 3% glutaraldehyde, 0.3% hydrogen peroxide in 0.1 mol l^{-1} 1,4-piperazine diethane sulfonic acid (PIPES) buffer (pH 7). After 2 h at room temperature, tissue was washed for 30 min in 0.1 mol l^{-1} PIPES buffer. Secondary fixation was achieved by immersion in 1% osmium tetroxide in PIPES buffer for 1 h at room temperature. After washing, specimens were dehydrated in graded ethanol and embedded in Araldite epoxy resin. Ultrathin sections (50 nm) were cut on a Reichert-Jung Ultracut S (Reichert-Jung). Sections were counterstained with uranyl acetate, followed by lead citrate, before viewing in a Philips CM100 electron microscope (Philips Electronics).

DNA extraction and quantification. The QIAamp DNA Blood Mini kit (Qiagen 51104) was used to extract genomic DNA from patients' blood for short tandem repeat analysis, HLA tissue typing and bisulfite sequencing. DNA was extracted from trophoblast organoids, decidual and placental tissues by digestion with ATL buffer (Qiagen 19076) and proteinase K (Sigma P4850), followed by purification steps with RNase A (Sigma R6513) and Protein Precipitation Solution (Qiagen 158910). DNA was precipitated with isopropanol and washed with 70% ethanol. DNA quality and concentration were determined in a Nanodrop ND-1000 Spectrophotometer.

Comparative genomic hybridization analysis. DNA from two independently derived trophoblast organoid samples at early and late passages was analysed with

© 2018 Springer Nature Limited. All rights reserved.

Agilent Sureprint G3 unrestricted CGH ISCA 8×60K array (Agilent G4450A). DNA samples from late passage organoids were compared to early passage organoids (hybridization control). DNA was diluted to 50 ng μl^{-1} and labelled using the Agilent kit following the manufacturer's instructions. Data analysis for segmentation and copy number calls was performed at a genome-wide resolution of 500 kb by the default analysis method, CGH v2 from the Agilent CytoGenomics software Edition 2.5.8.11 (Build 37).

Short tandem repeat analysis and HLA typing. Microsatellite analysis was performed with the GenePrint PowerPlex16 System (Promega) involving fluorescent labelled multiplexed PCR amplification of 15 short tandem repeat (STR) loci and amelogenin sex-determining fragments. PCR fragment size resolution was achieved with capillary electrophoresis on a 3730XL DNA Analyzer (Applied Biosystems) before analysis of the raw data and STR allele calling with GeneMapper Versions 4 and 5 (Life Technologies) fragment analysis sizing and genotyping software. All typing was performed blinded. The Promega PowerPlex 16 kit was designed for forensic testing and has a sensitivity that can detect down to 5% contamination. It is used for monitoring of post-haematopoietic stem-cell transplant chimaerism[25,26]. The DNA for HLA genotyping was processed via the workflows of the EFI accredited Clinical Histocompatibility Laboratory. Low resolution typing of the *HLA-A*, *HLA-B* and *HLA-C* genes was achieved with LABType kits (One Lambda), which rely on reaction patterns observed when sequence-specific DNA probes immobilized on fluorescent X-MAP polystyrene beads (Luminex) hybridize to biotin-labelled multiplexed gene-specific PCR amplicons. The hybrids were detected with a Liquichip 200 fluorimeter (QIAgen) and HLA allele assignment was from HLA Fusion software (One Lambda). Ultra-high resolution typing of HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1 and HLA-DPB1 was achieved with an 'in house' third generation sequencing pipeline using Pacific Biosciences' Single Molecule Real-Time DNA sequencing technology as previously described[27].

Bisulfite sequencing. Approximately 300 ng of DNA was used for bisulfite conversion performed with the EpiTect Bisulphite Kit (Qiagen 59110), according to the manufacturer's protocol. The *ELF5* promoter region was amplified as described previously[28]. PCR products were cloned and sequenced, confirming representation of distinct alleles.

Global DNA methylation analysis. Genomic DNA bisulfite was performed with the CEGX TrueMethyl kit (Cambridge Epigenetix /NuGEN) and used for microarray-based DNA methylation analysis, performed at GenomeScan (GenomeScan BV), on the HumanMethylation850 BeadChip (Illumina) and were scanned on the Illumina iScan system. The resulting iDAT files were imported and analysed by ChAMP (v2.9.10)[29,30]. Samples were processed and filtered for a probe detection *P*-value of < 0.01, probes with a bead count of less than 3 in at least 5% of samples, no CpG and known single nucleotide polymorphisms at probe starts, probes aligning to several locations, and quality control using the on array control probes[31]. Of the total probes on the array, 755,577 passed the filtering and quality-control steps. The BMIQ method was used to normalize the two probe types present on the array. Beta methylation values from the EPIC array range from 0 (unmethylated) to 1 (methylated) and are the equivalent of percentage methylation[32]. Genomic annotations were imported from FDb.InfiniumMethylation.hg19 and IlluminaHumanMethylationEPICmanifest[33]. Genomic features in Fig. 2g comprise the following numbers of assayed CpGs (in organoids, placenta and maternal blood, respectively): CpG islands, 48,799, 48,799 and 13,576; promoters, 151,270, 151,270 and 88,621; gene bodies, 280,284, 280,284 and 117,138; LINE1, 48,799, 48,799 and 13,576. Density plots are scaled to area. LINE1 elements were downloaded as tables from the UCSC Genome browser for hg19[34]. Maternal blood samples (normal) are taken from ArrayExpress accession E-GEOD-66210.

RNA extraction, quantification and quality control. Total RNA was isolated using the miRNeasy isolation kit (Qiagen 217004) with on-column DNase digestion (Qiagen 79254). Quantification of RNA was performed with the Quant-iT RiboGreen RNA Assay Kit (Thermo Fisher Scientific R11490) by measuring the intensity of fluorescence at 528 nm with a Synergy HT Multi-Mode Microplate Reader (BioTek Instruments) according to the manufacturer's instructions. RNA quality was assessed on the Agilent 2100 bioanalyzer (Thermo Fisher Scientific). RNA integrity number (RIN) of each tested sample was greater or equal to 8.

Reverse transcription and qPCR. The expression of *ELF5*, *ERVW-1* and *GCM1* was analysed with Taqman Gene expression assays (Applied Biosystems). Total RNA (500 ng–1 μg) was reverse transcribed with Superscript VILO Reverse Transcriptase (Thermo Fisher Scientific 11754050) in the presence of random hexamers and RNase inhibitor following the supplier's instructions. qPCR was performed on 7900HT Fast Real-Time PCR system (Applied Biosystems) as previously described[11]. Relative expression levels were normalized to the geometric mean of the three housekeeping genes *HPRT1*, *TOP1* and *TBP* using the $2^{-\Delta Ct}$ method. The expression of C19MC miRNAs *hsa-miR-517-5p*, *hsa-miR-517(a, b)-3p*, *hsa-miR-526b-3p*, *hsa-miR-525-3p* and reference gene *RNU48* was analysed by TaqMan miRNA assays (Applied Biosystems). Total RNA (10 ng) was reverse transcribed using miRNA-specific stem–loop reverse transcription primers and TaqMan microRNA reverse transcription kit (Applied Biosystems 4366596) according to the supplier's instructions. qPCR assays were run with qPCRBIO Probe Mix Lo-ROX (PCR Biosystems) containing specific probes on an Eppendorf Mastercycler RealPlex 2 instrument. C_t data were normalized to the *RNU48* internal control by the $2^{-\Delta Ct}$ method. All qPCR reactions included no-template controls and minus reverse transcriptase controls (–RT). For further details and for Taqman Assay IDs for each gene, see Supplementary Methods.

Microarray expression profiling and data analysis. The microarray experiment was performed at Cambridge Genomic Services at University of Cambridge with a species-specific Gene 2.1 ST Array Plate (Affymetrix) according to the manufacturer's instructions. In brief, 100 ng of total RNA was amplified for each sample with inline PolyA spike-in control and the WT PLUS amplification kit (Affymetrix). By using the in line hybridization controls, we successfully amplified samples with the GeneChip WT terminal labelling kit (Affymetrix). Plate arrays were processed on the GeneTitan instrument (Affymetrix) with the GeneTitan Hybridization, Wash and Stain kit (Affymetrix). Samples were hybridized to the array, washed, stained and scanned. CEL files generated were loaded in R using the oligo package from Bioconductor. The raw data were then processed after quality controls using the Robust Multichip Analysis method. The limma package (3.34.8) was used to make the comparisons, and results were corrected for multiple testing using the false discovery rate. Microarray probes without gene identifier (ensembl gene id) were filtered out. Initial quality control included principal component analysis and MDS plots. Finally, the quality of the data was assessed, and the correlation of the samples in the groups was compared. Heat maps were generated with the R package 'pheatmap' (1.0.8), which uses the Euclidean method. For the gene heat maps, the input is the normalized intensity matrix. GO terms enrichment was obtained with R package 'clusterProfiler' (3.6.0) with function 'enrichGO', and chord plots were generated with the R package 'GOplot' (1.0.2).

LC–MS/MS analysis of trophoblast organoid supernatants. Trophoblast organoids (day 10 after passaging) were grown in TOM and supernatants were collected after overnight incubation. The supernatants (500 μl) and an aliquot of growth media were acidified with 50 μl of 1% formic acid in water (v/v) and loaded directly onto an Oasis HLB Prime μ-elution 96 well SPE plate (Waters 186008052) and extracted as described previously[35]. The eluant was evaporated under oxygen-free nitrogen at 40 °C and the residue reduced and alkylated before overnight tryptic digestion. Protein digests were analysed using an Ultimate 3000 nano LC system coupled to a Q Exactive Plus Orbitrap mass spectrometer (ThermoScientific) as described previously[35]. The nano LC–MS/MS files obtained from the six different extracts were searched, combined and separately, using Peaks 8.5 software (BSI) against the human Swissprot database (downloaded on 26 October 2017). A tryptic digest setting was used and precursor and product ion tolerances were set at 10 p.p.m. and 0.05 Da respectively. The search parameters included a fixed modification of a carboxyamidomethylation on cysteine residues and variable modifications such as methionine oxidation, N-terminal pyro-glutamate, N-terminal acetylation and C-terminal amidation. A false discovery rate (FDR) value of 1% was applied at the peptide and a minimum of 1 unique peptide also was required. For further details, see Supplementary Methods.

Statistics and reproducibility. All experiments reported in this study have been reproduced with similar results derived from independent samples (tissues and organoids) from several patients. The number of times the experiments were repeated with independently derived trophoblast organoid cultures are reported in the figure legends and summarized in Supplementary Table 1c. Given the descriptive nature of the work and biological variation between human samples, the experimental data points for each patient sample are shown separately unless stated otherwise. Trophoblast organoid culture protocols were independently replicated by four scientists. Statistical analyses used to analyse microarray, methylation and LC–MS/MS data are reported above. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment unless stated otherwise (for example, see 'short tandem repeat analysis and HLA typing' section).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Code used to analyse microarray data and EPIC array samples is available at https://github.com/CTR-BFX/2018-Turco-Moffett.

Data availability

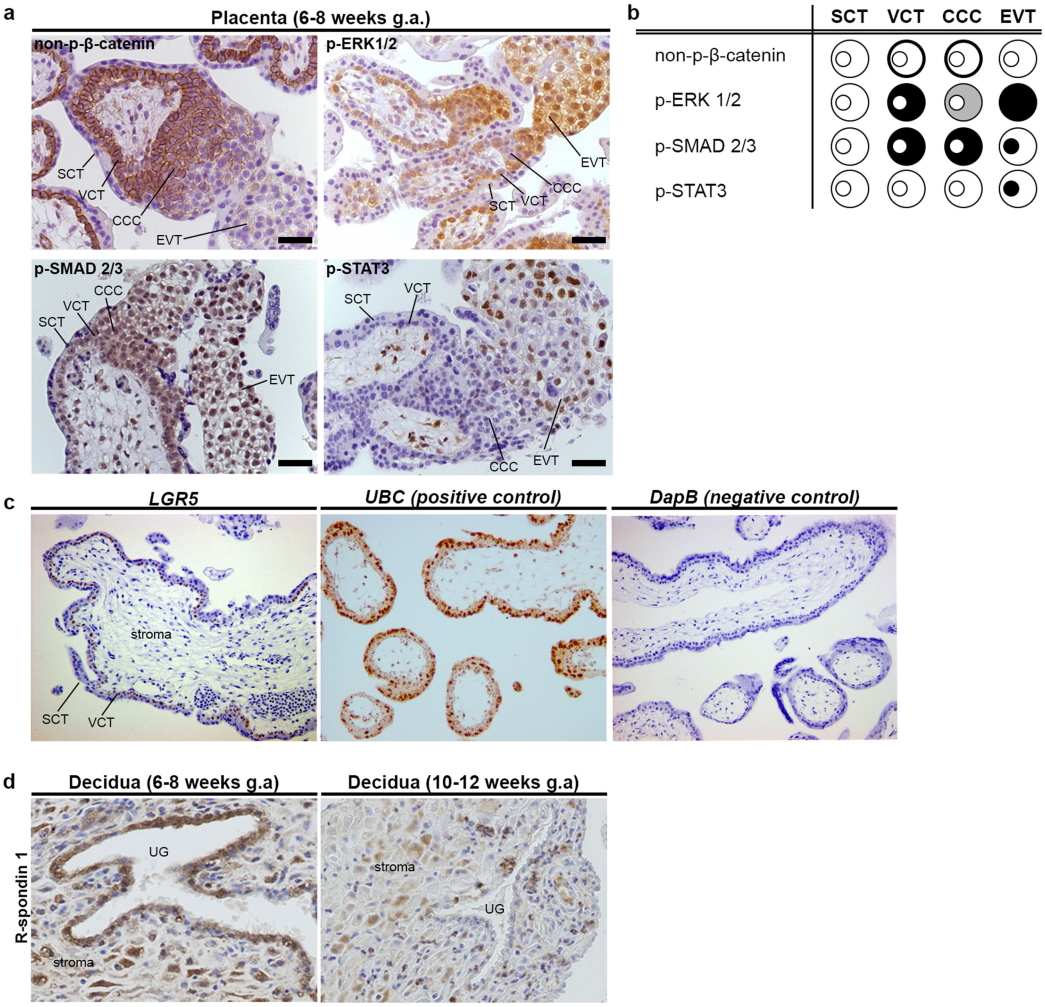
Microarray data for Fig. 2 and Extended Data Figs. 4 and 5 have been deposited in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-6683. Illumina EPIC methylation array data for Fig. 2 and Extended Data Fig. 5 have been deposited in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-7204. The mass spectrometry proteomics data for Extended Data Table 3 and Supplementary Table 2 have been deposited to the ProteomeXchange Consortium via the PRIDE[36] partner repository (https://www.ebi.ac.uk/pride/archive/) with the dataset identifier PXD009118 and 10.6019/PXD009118. All

© 2018 Springer Nature Limited. All rights reserved.

other data that support the findings of this study are available from the corresponding authors upon reasonable request.

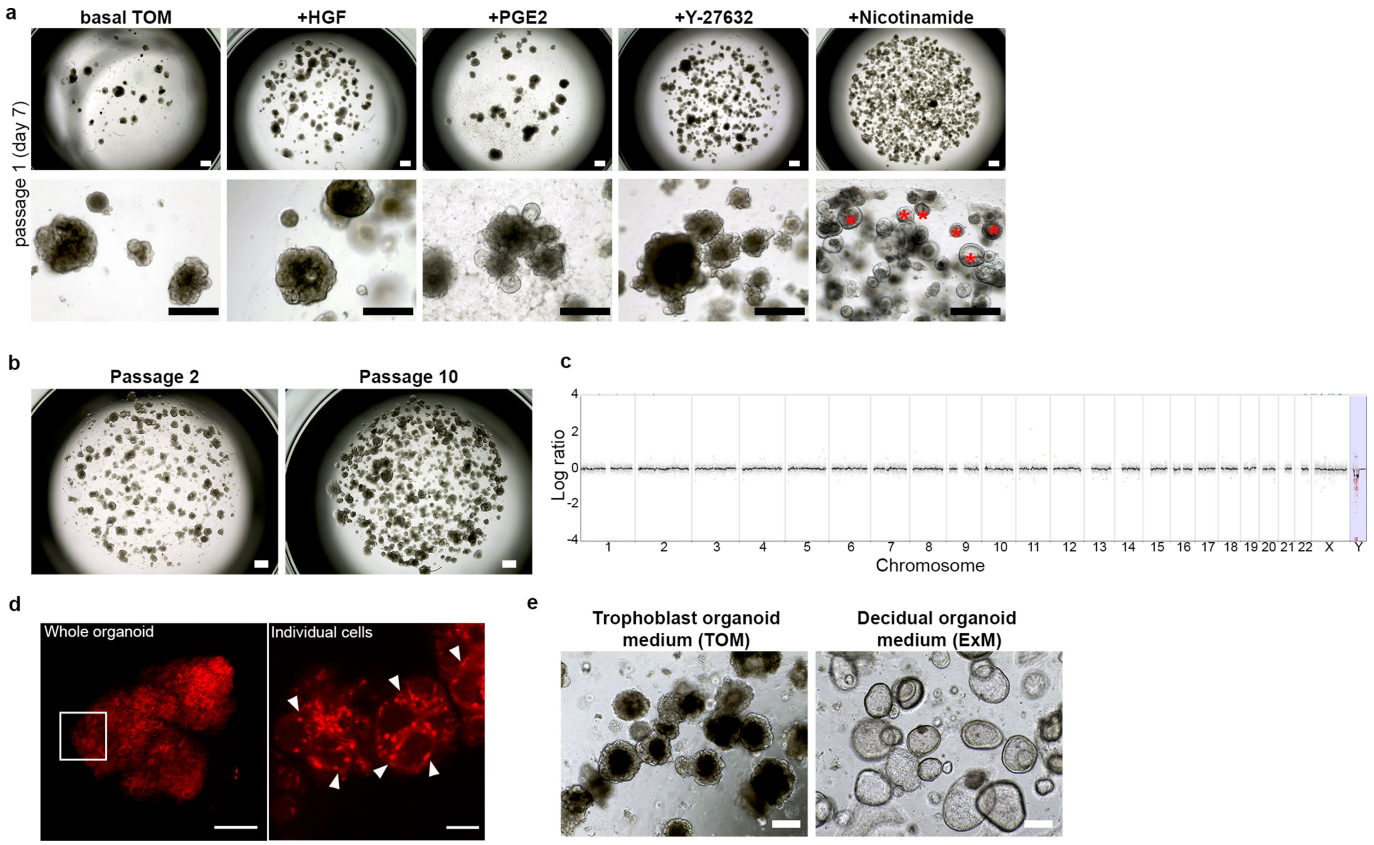
24. Turco, M. Y. et al. Establishment and differentiation of long-term trophoblast organoid cultures from the human placenta. *Protoc. Exchange* https://doi.org/10.1038/protex.2018.121 (2018).
25. Nollet, F., Billiet, J., Selleslag, D. & Criel, A. Standardisation of multiplex fluorescent short tandem repeat analysis for chimerism testing. *Bone Marrow Transplant.* **28**, 511–518 (2001).
26. Mossallam, G. I., Smith, A. G. & McFarland, C. Comparison of variable number tandem repeat and short tandem repeat genetic markers for qualitative and quantitative chimerism analysis post allogeneic stem cell transplantation. *J. Egypt. Natl. Canc. Inst.* **17**, 103–113 (2005).
27. Turner, T. R. et al. Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines. *HLA* **91**, 88–101 (2018).
28. Hemberger, M., Udayashankar, R., Tesar, P., Moore, H. & Burton, G. J. ELF5-enforced transcriptional networks define an epigenetically regulated trophoblast stem cell compartment in the human placenta. *Hum. Mol. Genet.* **19**, 2456–2467 (2010).
29. Morris, T. J. et al. ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428–430 (2014).
30. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
31. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
32. Teschendorff, A. E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).
33. Fortin, J.-P., Triche, T. J. Jr & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).
34. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
35. Kay, R. G., Galvin, S., Larraufie, P., Reimann, F. & Gribble, F. M. Liquid chromatography/mass spectrometry based detection and semi-quantitative analysis of INSL5 in human and murine tissues. *Rapid Commun. Mass Spectrom.* **31**, 1963–1973 (2017).
36. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, 11033 (2016).
37. Zhang, J. et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111.010587 (2012).

© 2018 Springer Nature Limited. All rights reserved.



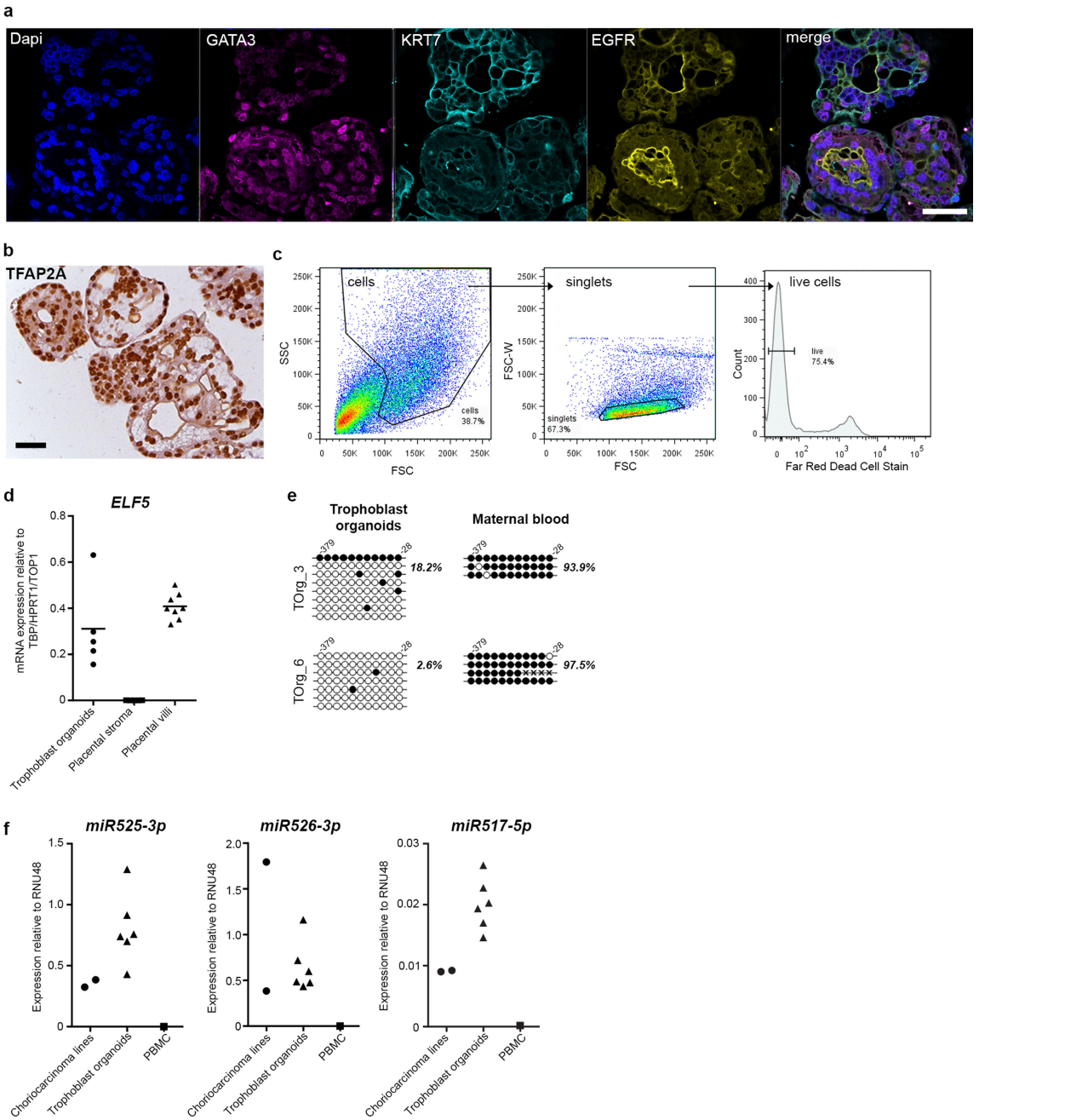
Extended Data Fig. 1 | Staining for signalling pathways in first trimester placenta and decidua. a, Immunohistochemical staining of first trimester placenta (6–8 weeks gestation) for effectors of the following major signalling pathways: (1) WNT signalling through non-phosphorylated-β-catenin (S33/S37/T41); (2) TGFβ signalling through phosphorylated (p)-SMAD2 (S465/467) and p-SMAD3 (S423/425); (3) MAPK signalling through p-ERK1 and p-ERK2 (T202/Y204); and (4) p-STAT3 (Y705) signalling. Scale bars, 50 μm. Representative images from $n = 8$ for each antibody. BMP signalling through SMAD1, SMAD5 or SMAD8 was not possible to assess by immunohistochemistry. VCT cells and CCCs displayed membrane-localized staining of non-phosphorylated-β-catenin, whereas p-ERK1 and p-ERK2 was mostly found in the cytoplasm in both cell types. Cytoplasmic and nuclear signals for p-ERK1 and p-ERK2 were detected in the EVT. p-SMAD2 and p-SMAD3 staining also showed stronger nuclear signals in the EVT, suggesting a role for TGFβ signalling in differentiation, in accordance with a previous report[9]. Phosphorylated nuclear STAT3 was detected only in the EVT, again indicating involvement in their differentiation. The SCT was negative for all these signals. **b**, Summary of findings from **a**. Trophoblast cells from different regions of the placenta are represented as a circle with a nucleus (small inner circle). Black indicates strong staining, grey indicates faint staining, and white denotes not detected. Thicker circles indicate staining localized to the cell membrane. **c**, In situ hybridization for *LGR5* on first trimester placental villi. *LGR5* transcripts are detected in the VCT. Stroma is negative. Positive-control probe is for *UBC*; negative-control probe is for the bacterial gene *dapB*. Nuclei are counterstained with haematoxylin. Images are at ×10 magnification. Representative images from $n = 2$. **d**, Immunohistochemistry staining for R-spondin 1 in early first trimester (6–8 weeks gestation) and late first trimester (10–12 weeks gestation) decidual samples. Representative images from $n = 2$ for each tissue type. Images are at ×20 magnification. UG, uterine glands.

© 2018 Springer Nature Limited. All rights reserved.



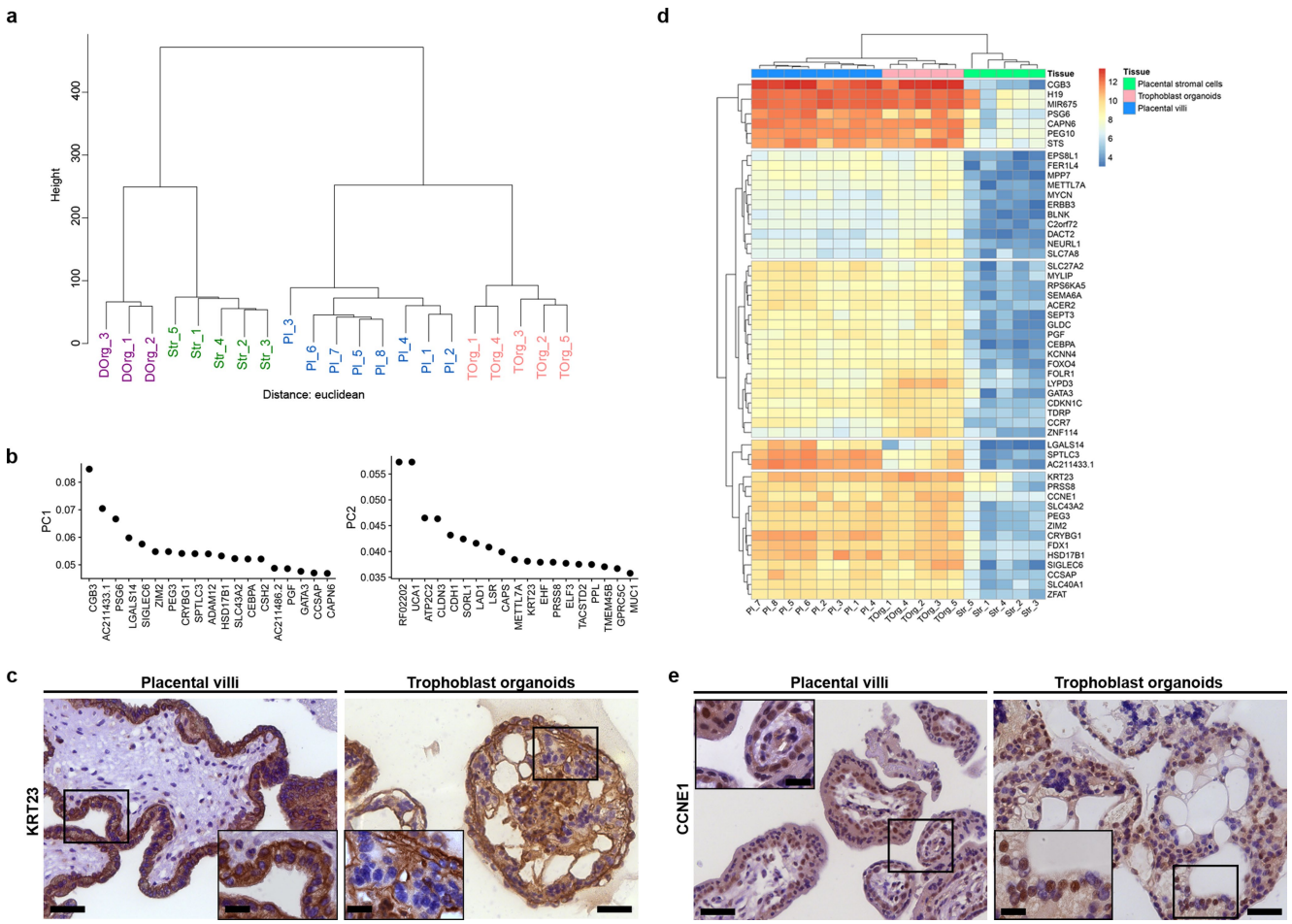
Extended Data Fig. 2 | Culture components tested for the establishment of long-term organoid cultures of trophoblast from human placentas. **a**, Growth factors (HGF, PGE2, Y-27632 and nicotinamide) were added as supplements to basal TOM that contains EGF, CHIR99021, R-spondin 1, A83-01 and FGF2 (Supplementary Table 1a, b). Bright-field images of placental digests at passage 1, day 7. The cystic structures that appear with the addition of nicotinamide (red asterisks) are contaminating maternal glandular organoids. Representative images from $n = 2$. Conditions containing factors that did not show growth are not included. Scale bars, 500 μm. **b**, Trophoblast organoid cultures at passages 2 and 10 with continuous culture. Representative images from $n = 3$. Scale bars, 500 μm. **c**, Analysis of genetic stability of cultures ($n = 2$) with comparative genomics hybridization (CGH) array. Shown is a representative whole-genome array CGH plot generated with Agilent Cytogenomics software. Genomic DNA from late passage (passage 8) trophoblast organoids is compared to genomic DNA from an early passage (passage 2). Each spot is a single probe. Plotted are the log ratios of the average signal intensity of each probe on the *y* axis along its position on the chromosomes (1–22, X and Y) on the *x* axis. A log signal ratio of 0 represents equivalent copy number in the samples. No significant DNA copy number abnormalities were identified. **d**, Live imaging of trophoblast organoid cultures ($n = 2$) passaged for more than 6 months and then frozen, thawed and exposed to Mitotracker Red. Functional mitochondria are visible showing that the cells are healthy (white arrowheads). Scale bars, 50 μm (whole organoid) and 10 μm (individual cells). **e**, Organoids derived from the same placental cell isolate using either trophoblast or decidual organoid medium (TOM or ExM, respectively) demonstrate that matched placental (fetal) and decidual (maternal) organoids can be derived from one sample. Representative bright-field images from $n = 3$. Scale bars, 500 μm.

© 2018 Springer Nature Limited. All rights reserved.



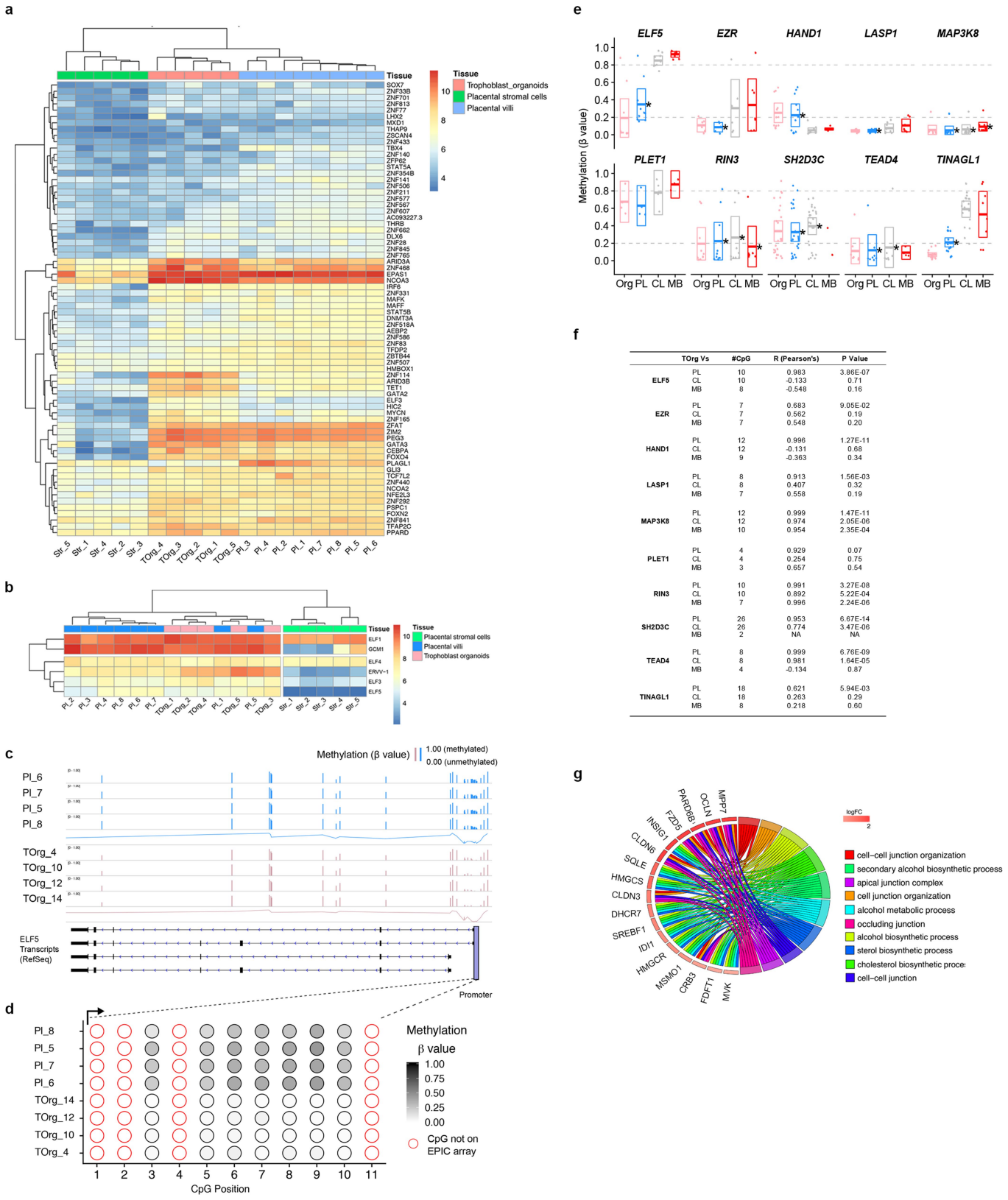
Extended Data Fig. 3 | Trophoblast organoids retain characteristic features of first trimester trophoblast in vivo. a, Immunofluorescence images by confocal microscopy of GATA3, KRT7, EGFR and DAPI on trophoblast organoids (representative images from $n = 3$). EGFR stains both VCT cells and the surface of the SCT, as reported in vivo[7]. The basement membrane is around the outside of the organoids, with syncytial masses present in the centre. Scale bar, 50 μm. **b**, Immunohistochemical staining for the transcription factor TFAP2A shows uniform expression on trophoblast organoids (representative image from $n = 20$). Scale bars, 50 μm. **c**, Gating strategy used for flow cytometric analysis of single, live cells from trophoblast organoids. **d**, qPCR analysis of *ELF5* in trophoblast organoids ($n = 5$) compared to whole placental villi ($n = 8$) and placental stromal cells ($n = 5$). Graph shows expression levels relative to the geometric mean of the three housekeeping genes *TBP*, *TOP1* and *HPRT1*. The mean *ELF5* expression is shown for each sample group. **e**, Bisulfite sequencing of the *ELF5* promoter region (−379 bp to −28 bp upstream of the transcription start site) of trophoblast organoids from two different placentas (TOrg_3 and TOrg_6) and matched maternal blood leukocytes (positive control). The relative percentage of methylated cytosine residues (filled circles) is indicated. **f**, qPCR analysis for miRNAs *miR525-3p*, *miR526-3p* and *miR517-5p* from the C19MC cluster on trophoblast organoids ($n = 6$), choriocarcinoma lines JEG-3 and JAR cells (positive controls) and PBMC (low expression/negative control). Graph shows relative expression levels of each organoid culture to the housekeeping gene *RNU48*.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Hierarchical clustering of microarray data comparing placental villi, trophoblast organoids and placental stromal cells. a, Unsupervised hierarchical clustering analysis of global gene expression profiles by microarray of first trimester placental villi (Pl) ($n = 8$), trophoblast organoids (TOrg) ($n = 5$), placental stromal cells (Str) ($n = 5$) and decidual organoids (DOrg) ($n = 3$). Analysis was based on 12,673 probes. The expression profiles of trophoblast organoids cluster with first trimester placental samples, whereas decidual organoids and placental stromal cells cluster in a separate tree. **b**, The top 20 genes contributing to PC1 and PC2 in the principal component analysis plot from Fig. 2e. The top genes contributing to PC1 are trophoblast-specific genes such as *CGB3*, *GATA3* and *PSG6*, indicating that these genes separate the trophoblast organoid and placental villous samples from the two potentially contaminating, non-trophoblast samples (decidual organoids and placental stromal cells). The top genes contributing to PC2 are epithelial genes such as *CLDN3*, *TACSTD2* and *KRT23*. The organoids only contain trophoblast, but cells of the villous core (stromal, Hofbauer and endothelial cells) are also present in the placental samples. **c**, Immunohistochemistry of placental villi and trophoblast organoids stained for KRT23 shows expression in all trophoblast cells in vivo and in vitro. The experiment was repeated independently three times. Scale bars, 50 μm and 20 μm (insets). **d**, Clustered heat map of differentially expressed genes between trophoblast organoids, placental villi and placental stroma with an absolute base-2 logarithmic fold change of 2 (adjusted $P < 0.05$). **e**, Immunohistochemistry of placental villi and trophoblast organoids stained for CCNE1 shows expression in trophoblast cells in vivo and in vitro. Scale bars, 50 μm and 20 μm (insets). The experiment was repeated independently three times.

© 2018 Springer Nature Limited. All rights reserved.



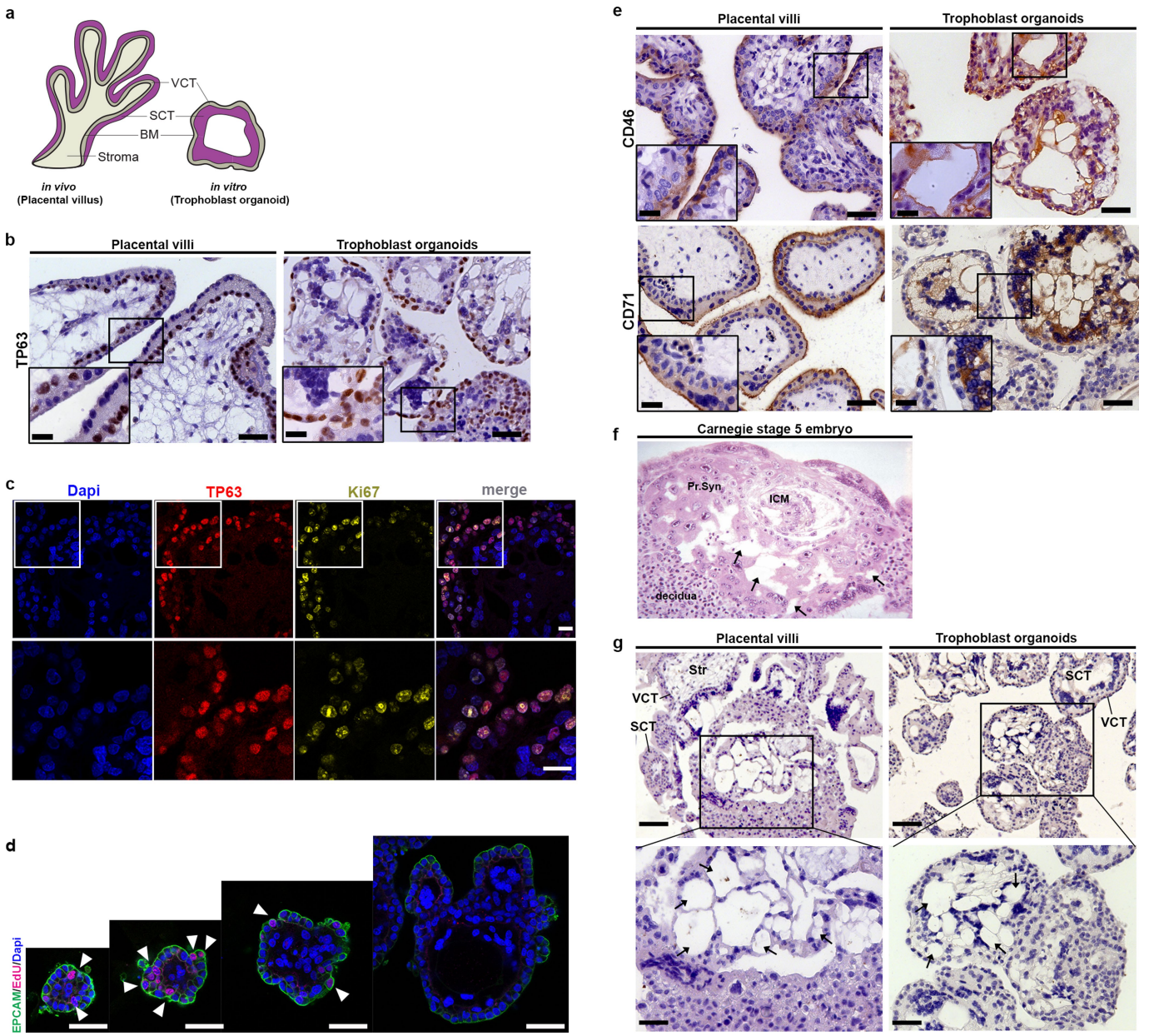
	TOrg Vs	#CpG	R (Pearson's)	P Value
ELF5	PL	10	0.983	3.86E-07
	CL	10	-0.133	0.71
	MB	8	-0.548	0.16
EZR	PL	7	0.683	9.05E-02
	CL	7	0.562	0.19
	MB	7	0.548	0.20
HAND1	PL	12	0.996	1.27E-11
	CL	12	-0.131	0.68
	MB	9	-0.363	0.34
LASP1	PL	8	0.913	1.56E-03
	CL	8	0.407	0.32
	MB	7	0.558	0.19
MAP3K8	PL	12	0.999	1.47E-11
	CL	12	0.974	2.05E-06
	MB	10	0.954	2.35E-04
PLET1	PL	4	0.929	0.07
	CL	4	0.254	0.75
	MB	3	0.657	0.54
RIN3	PL	10	0.991	3.27E-08
	CL	10	0.892	5.22E-04
	MB	7	0.996	2.24E-06
SH2D3C	PL	26	0.953	6.67E-14
	CL	26	0.774	3.47E-06
	MB	2	NA	NA
TEAD4	PL	8	0.999	6.76E-09
	CL	8	0.981	1.64E-05
	MB	4	-0.134	0.87
TINAGL1	PL	18	0.621	5.94E-03
	CL	18	0.263	0.29
	MB	8	0.218	0.60

Extended Data Fig. 5 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

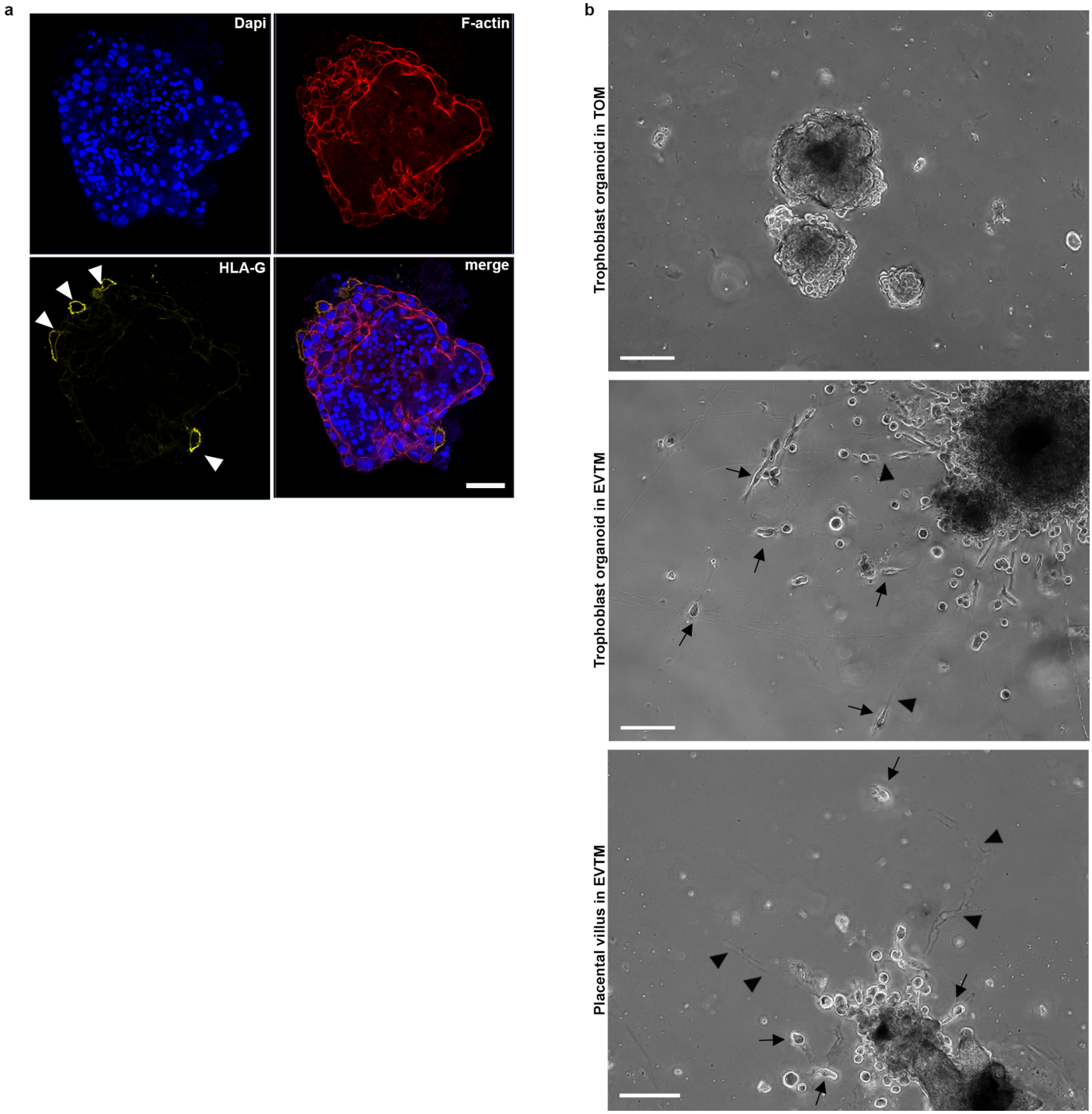
Extended Data Fig. 5 | Expression profiles of transcription factors in trophoblast organoids and placental villi. a, Heat map highlighting transcription factors from the differentially expressed genes between placental villi, trophoblast organoids and placental stromal cells. **b**, Heat map of genes from the *ELF* family of transcription factors and syncytial genes, *GCM1* and *ERVW-1*. *ELF3* and *ELF5* both show moderate expression levels across the organoids and placental samples, and very low or no expression in the stromal samples. *ELF4* and *ELF1* are similar in all samples. There is very high expression of *ELF1* in placentas and organoids. Similarly, both *ERVW-1* and *GCM1* are expressed at higher levels in placentas and organoids in agreement with qPCR data (Fig. 3d). **c**, Genomic mapping of the methylation array probes to the *ELF5* gene. The height of the bars indicates the methylation level from 0 (unmethylated) to 1.0 (fully methylated). **d**, Methylation of the *ELF5* promoter shows hypomethylation in the organoid and placental samples. **e**, Distribution of methylcytosine across the promoters of the 10 trophoblast gatekeeper genes identified previously in mice[12]. The organoids (Org) and placental (Pl) samples show very similar methylation patterns across all 10 gene promoters that are distinct from the control brain (CL) and maternal blood (MB). Box plots comprise minimum: 1.5× interquartile range, bottom: first quartile, middle: median, top: third quartile, maximum: 1.5× interquartile range. *$P < 0.01$, Pearson's correlation coefficient. **f**, Table for results in **e** showing Pearson's correlation coefficient (R), number of CpG islands or probes compared and P values. **g**, Chord plot representing terms from the Gene Ontology analysis of upregulated genes in trophoblast organoids.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Trophoblast organoids proliferate and structurally resemble first trimester placental villi. a, A schematic diagram of a normal placental villus in vivo compared to a trophoblast organoid. The basement membrane (BM) beneath the VCT is contiguous with the stromal villous core in vivo and with the Matrigel in vitro. The SCT contacts maternal blood in the intervillous space in vivo, and forms in the centre of the organoids. **b**, Immunohistochemical staining for TP63 in first trimester placenta and trophoblast organoids (representative images from $n = 14$). TP63 is expressed in VCT. Scale bars, 50 μm and 20 μm (insets). **c**, Representative images of TP63, Ki67 and DAPI staining in trophoblast organoids by confocal microscopy ($n = 3$). Cells on the outside of the organoids are $TP63^+$ and $Ki67^+$. Scale bars, 20 μm. **d**, Confocal microscopy images of trophoblast organoid stained for EdU, EPCAM and DAPI, showing fewer proliferating cells (white arrowheads) as the organoids enlarge. Scale bars, 50 μm. Representative images from $n = 3$. **e**, Immunohistochemical staining for the SCT markers CD46 and CD71 in first trimester placenta and trophoblast organoids (representative images from $n = 20$). CD46 and CD71 stain the syncytial brush border. Scale bars, 50 μm and 20 μm (insets). **f**, Carnegie stage 5b embryo (approximately 9 days after fertilization) from the Carnegie Collection at the early lacunar stage (number 8171). Courtesy of A. Enders and the Centre for Trophoblast Research (https://www.trophoblast.cam.ac.uk/Resources/enders). Arrows point to examples of cavities that appear in the primitive syncytium owing to fluid accumulation before the coelomic cavity and the embryo have fully developed. **g**, Similar cavities in placental tissue samples from first trimester (6–9 weeks gestation) and in syncytium in the centre of trophoblast organoids. Boxed areas are shown at higher magnification (bottom). Scale bars, 200 μm (top) and 50 μm (bottom). Similar morphology seen in at least five early placental villi and in all organoids. ICM, inner cell mass; Pr.Syn., primitive syncytium; Str, stromal core.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Trophoblast organoids grown in TOM and EVTM. a, Confocal image of organoid stained for F-actin, DAPI and HLA-G with a merged image. A few isolated cells stain for HLA-G (arrowheads) at the periphery of the organoid. Scale bar, 50 μm. Representative images from $n = 3$. **b**, Phase-contrast images from time-lapse videos of trophoblast organoids grown for 16 h in TOM when EVT differentiation does not occur (top) and no invasive cells are visible. An organoid (middle) and a placental villous explant (bottom) exposed to EVTM are shown for comparison. Arrows indicate cells migrating, and arrowheads denote the visible tracks made as the cells invade through the Matrigel. For time-lapse videos of these cultures, see Supplementary Videos 1, 4 and 5. Scale bars, 200 μm.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Microsatellite analysis and HLA typing of organoids from placental digests in TOM show that they are of fetal origin

a

Sample ID	D3S1358	TH01	D21S11	D18S51	Penta E	D5S818	D13S317	D7S820	D16S539	CSF1PO	Penta D	Amelo.	vWA	D8S1179	FGA
Maternal Blood_2	16,17	7,9.3	28,30	12,14	15,15	12,13	10,15	11,11	12,13	10,10	9,9	x,x	15,18	12,13	22.2,23
Decidua_2	16,17	7,9.3	28,30	12,14	15,15	12,13	10,15	11,11	12,13	10,10	9,9	x,x	15,18	12,13	22.2,23
Placenta_2	14,16	6,9.3	28,30	14,15	12,15	11,13	10,14	10,11	12,12	10,10	9,13	x,x	15,17	13,13	22,23
Trophoblast Organoid_2	14,16	6,9.3	28,30	14,15	12,15	11,13	10,14	10,11	12,12	10,10	9,13	x,x	15,17	13,13	22,23
Maternal Blood_3	16,17	6,9	29,31.2	12,14	7,12	11,12	10,11	9,10	11,11	10,12	10,13	x,x	18,18	12,13	20,25
Decidua_3	16,17	6,9 (9.3)	29,31.2 (28)	12,14	7,12 (19)	11,12	10,11 (14)	9,10	11,11	10,12 (11)	10,13 (14)	x,x	18,18 (19)	12,13	20,25 (21)
Placenta_3	16,16	9,9.3	28,31.2	12,14	7,19	11,11	10,14	9,10	11,11	11,12	13,14	x,x	18,19	12,12	21,25
Trophoblast Organoid_3	16,16	9,9.3	28,31.2	12,14	7,19	11,11	10,14	9,10	11,11	11,12	13,14	x,x	18,19	12,12	21,25
Maternal Blood_6	14,15	7,9.3	30,30	12,16	16,19	11,11	11,13	9,12	11,13	10,11	9,9	x,x	15,15	10,10	21,26
Decidua_6	14,15	7,9.3	30,30	12,16	16,19	11,11	11,13	9,12	11,13	10,11	9,9	x,x	15,15	10,10	21,26
Placenta_6	15,17	6,7	28,30	16,17	5,16	10,11	12,13	9,11	11,13	10,10	9,13	x,y	15,17	10,11	24,26
Trophoblast Organoid_6	15,17	6,7	28,30	16,17	5,16	10,11	12,13	9,11	11,13	10,10	9,13	x,y	15,17	10,11	24,26
Maternal Blood_7	15,16	6,6	29,30.2	16,16	7,18	9,13	12,12	9,10	13,13	11,11	10,11	x,x	17,18	12,14	21,21
Decidua_7	15,16 (17)	6,6	29,30.2	16,16	7,18	9,13	12,12 (10)	9,10	13,13 (10)	11,11	10,11	x,x (y)	17,18	12,14 (13)	21,21 (24)
Placenta_7	15,17	6,6	29,29	16,16	7,18	13,13	10,12	9,10	10,13	11,11	10,11	x,y	17,18	13,14	21,24
Trophoblast Organoid_7	15,17	6,6	29,29	16,16	7,18	13,13	10,12	9,10	10,13	11,11	10,11	x,y	17,18	13,14	21,24
Maternal Blood_8	17,18	7,9	30,30	16,20	11,12	12,12	11,12	9,10	11,14	11,12	11,13	x,x	15,18	13,14	22,22
Decidua_8	17,18 (15)	7,9 (8)	30,30	16,20 (12)	11,12 (7)	12,12 (10)	11,12	9,10 (11)	11,14	11,12	11,13	x,x (y)	15,18	13,14	22,22 (28)
Placenta_8	15,18	7,8	30,30	12,16	7,12	10,12	11,11	9,11	9,14	11,11	11,11	x,y	15,17	13,14	22,28
Trophoblast Organoid_8	15,18	7,8	30,30	12,16	7,12	10,12	11,11	9,11	9,14	11,11	11,11	x,y	15,17	13,14	22,28
Maternal Blood_9	15,18	9.3,9.3	33.2,34.2	12,17	5,5	12,13	12,12	10,11	11,11	10,12	11,12	x,x	15,16	14,15	19,22
Decidua_9	15,18	9.3,9.3 (10)	33.2,34.2 (31)	12,17 (13)	5,5 (14)	12,13	12,12	10,11(8)	11,11	10,12	11,12 (10)	x,x (y)	15,16	14,15 (12)	19,22 (21)
Placenta_9	15,15	9.3,10	31,33.2	13,17	5,14	12,13	12,12	8,10	11,11	10,12	10,12	x,y	16,16	12,14	19,21
Trophoblast Organoid_9	15,15	9.3,10	31,33.2	13,17	5,14	12,13	12,12	8,10	11,11	10,12	10,12	x,y	16,16	12,14	19,21

b

Sample ID	HLA-A*	HLA-A*	HLA-B*	HLA-B*	HLA-C*	HLA-C*	HLA-DRB1*	HLA-DRB1*	HLA-DQB1*	HLA-DQB1*	HLA-DPB1*	HLA-DPB1*
Maternal Blood_2	02:01:01:01/16	02:01:01:01/16	35:01:01:02/04/05/06	44:02:01:03	07:04:01:01/03	03:03:01:01	11:01:01	08:01/77	03:01:01	04:02:01	04:01:01	04:01:01
Trophoblast Organoid_2	02:01:01:01/16	11:01:01:01	35:01:01:02/04/05/06	39:01:01:03/05	12:03:01:01	03:03:01:01	11:01:01	08:01/77	03:01:01	04:02:01	04:02:01	04:01:01
Maternal Blood_3	02:01:01:01/16	24:02:01:01	15:01:01:01	40:01:02:01/04	03:04:01:01	03:03:01:01	04:04:01	01:01:01	03:02:01	05:01:01	04:01:01	16:01:01
Trophoblast Organoid_3	02:01:01:01/16	24:02:01:01	15:01:01:01	15:01:01:01	03:04:01:01	03:03:01:01	04:04:01	04:01:01	03:02:01	03:02:01	04:01:01	04:01:01
Maternal Blood_6	03:01:01:01	30:02:01:01	07:02:01:01/03	18:01:01:01/06	05:01:01:01	07:02:01:03	01:03:01	01:03:01	03:01:01	05:01:01	02:01:02/19	04:01:01
Trophoblast Organoid_6	68:01:02:02	30:02:01:01	27:05:02:01	18:01:01:01/06	05:01:01:01	07:04:01:01/03	01:03:01	08:01/77	04:02:01	05:01:01	02:01:02/19	04:01:01
Maternal Blood_7	32:01:01:01	33:01:01:01	14:02:01:01	44:02:01:01	05:01:01:02	08:02:01:01	01:01:01	01:02:01	05:01:01	05:01:01	13:01/107:01	04:01:01
Trophoblast Organoid_7	24:02:01:04	33:01:01:01	14:02:01:01	44:02:01:01	05:01:01:02	08:02:01:01	13:01:01	01:02:01	05:01:01	06:03:01	02:01:02/19	04:01:01
Maternal Blood_8	24:02:01:01	25:01:01:01	18:01:01:02/05	27:05:02:01	01:02:01:01	12:03:01:01	15:01:01	13:03:01	03:01:01	06:02:01	04:01:01	03:01:01
Trophoblast Organoid_8	24:02:01:01	02:01:01:01/16	49:01:01	27:05:02:01	01:02:01:01	07:02:01:01/15	11:02:01	13:03:01	03:01:01	03:19:01	10:01:01	03:01:01
Maternal Blood_9	26:01:01:01	03:01:01:01	07:02:01:01/03	27:05:02:01	02:02:01:01	07:02:01:03	04:04:01	04:04:01	03:02:01	03:02:01	04:01:01	02:01:02/19
Trophoblast Organoid_9	26:01:01:01	02:01:01:01/16	44:02:01:01	27:05:02:01	02:02:01:01	05:01:01:02	04:01:01	04:04:01	03:01:01	03:02:01	04:02:01	02:01:02/19

a, PowerPlex16 short tandem repeat (STR) genotyping of DNA from matched maternal and fetal samples (maternal blood, decidua and placenta) to identify origin of organoids. The 15 STR loci analysed are shown. Numbered STR alleles observed for each DNA at a particular locus are listed within the relevant column. As expected, in most cases a maximum of two alleles was seen. In cases where there was evidence of an additional allele from a fetal-derived STR haplotype, the allele number appears in brackets. This is consistent with the decidua containing fetal extravillous trophoblast. The results at informative loci in which the fetal and trophoblast organoids match are underlined. **b**, HLA genotyping with third-generation single-molecule real-time sequencing of DNA from matched maternal blood and trophoblast organoid to confirm STR analysis. HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1 and HLA-DPB1 loci were investigated. The two HLA alleles at each locus defined at high resolution are shown for each sample. For each locus, the fetal trophoblast will share one allele with the mother but the other allele will be derived from the father and is likely to be different. Not all loci are informative as some paternal and non-inherited maternal alleles are the same.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Microsatellite analysis and HLA typing of organoids from placental digests derived in ExM show that they are of maternal origin

a

Sample ID	D3S1358	TH01	D21S11	D18S51	Penta E	D5S818	D13S317	D7S820	D16S539	CSF1PO	Penta D	Amelo.	vWA	D8S1179	FGA
Maternal Blood_1	15, 15	7, 9	30, 30	14, 17	11, 12	12, 13	11, 12	8, 8	11, 12	12, 12	9, 12	xx	17, 19	13, 15	22, 25
Decidua_1	15, 15	7, 9	30, 30	14, 17	11, 12	12, 13	11, 12	8, 8	11, 12	12, 12	9, 12	xx	17, 19	13, 15	22, 25
Placenta_1	15, 15	7, 8	29, 30	17, 17	10, 12	11, 12	11, 12	8, 10	12, 12	10, 12	10, 12	xx	16, 17	13, 15	22, 25
Decidual Organoid_1	15, 15	7, 9	30, 30	14, 17	11, 12	12, 13	11, 12	8, 8	11, 12	12, 12	9, 12	xx	17, 19	13, 15	22, 25
Maternal Blood_2	16, 17	6, 9.3	28, 28	15, 19	12, 12	11, 12	8, 11	12, 14	12, 12	10, 11	11, 12	xx	16, 17	12, 15	20, 21
Decidua_2	16, 17	6, 9.3	28, 28	15, 19	12, 12	11, 12	8, 11	12, 14	12, 12	10, 11	11, 12	xx	16, 17	12, 15	20, 21
Placenta_2	16, 17	6, 8	28, 31.2	15, 19	12, 15	11, 12	8, 9	10, 12	12, 13	10, 11	11, 12	xx	16, 17	10, 15	20, 23
Decidual Organoid_2	16, 17	6, 9.3	28, 28	15, 19	12, 12	11, 12	8, 11	12, 14	12, 12	10, 11	11, 12	xx	16, 17	12, 15	20, 21
Maternal Blood_3	14, 17	6, 6	28, 31.2	12, 12	7, 13	12, 12	11, 12	10, 12	11, 13	10, 11	9, 10	xx	15, 17	13, 13	23, 24
Decidua_3	14, 17	6, 6	28, 31.2	12, 12, (15)	7, 13	12, 12, (13)	11, 12	10, 12, (9)	11, 13	10, 11, (12)	9, 10, (13)	xx (y)	15, 17	13, 13, (14)	23, 24, (22)
Placenta_3	17, 17	6, 6	28, 28	12, 15	7, 12	12, 13	11, 11	9, 10	13, 13	11, 12	10, 13	xy	15, 17	13, 14	22, 24
Decidual Organoid_3	14, 17	6, 6	28, 31.2	12, 12	7, 13	12, 12	11, 12	10, 12	11, 13	10, 11	9, 10	xx	15, 17	13, 13	23, 24
Maternal Blood_4	15, 15	9.3, 9.3	30, 32.2	18, 22	12, 14	11, 12	8, 12	10, 12	12, 13	9, 10	12, 16	xx	15, 16	13, 13	19, 22
Decidua_4	15, 15	9.3, 9.3	30, 32.2	18, 22	12, 14	11, 12	8, 12	10, 12	12, 13	9, 10	12, 16	xx	15, 16	13, 13	19, 22
Placenta_4	15, 16	7, 9.3	30, 30	14, 18	11, 12	11, 12	12, 14	9, 12	13, 14	10, 10	9, 16	xx	16, 16	13, 14	20, 22
Decidual Organoid_4	15, 15	9.3, 9.3	30, 32.2	18, 22	12, 14	11, 12	8, 12	10, 12	12, 13	9, 10	12, 16	xx	15, 16	13, 13	19, 22

b

Sample ID	HLA-A	HLA-B	HLA-C
Maternal Blood_1	01,02	41,52	12,17
Decidua_1	01,02	41,52	12,17
Placenta_1	01,02	27,52	01,12
Decidual Organoid_1	*Fail x2*	41,52	12,17
Maternal Blood_2	02,02	07,44	05,07
Decidua_2	02,02	07,44	05,07
Placenta_2	02,02	07,44	05,07
Decidual Organoid_2	02,02	07,44	05,07
Maternal Blood_3	25,29	18,44	12,16
Decidua_3	*Fail x2*	18,44	12,16
Placenta_3	23,25	18,44	04,12
Decidual Organoid_3	25,29	18,44	12,16
Maternal Blood_4	23,24	13,44	04,04
Decidua_4	23,24	13,44	04,04
Placenta_4	11,24	13,40	02,04
Decidual Organoid_4	23,24	13,44	04,04

a, PowerPlex16 STR genotyping of DNA from matched maternal and fetal samples (decidua, maternal blood and placenta) to identify origin of organoids. The 15 STR loci analysed are shown. Numbered STR alleles observed for each DNA at a particular locus are listed within the relevant column. As expected, in most cases a maximum of two alleles was seen, but where there was lesser evidence of an additional allele from a fetal derived STR haplotype the allele number appears in brackets. The results at informative loci in which maternal blood, decidua and decidual organoids match are underlined. **b**, HLA genotyping with LABType reverse sequence-specific oligonucleotide probes (SSOP) of DNA from four pregnancies with matched decidua, blood, placenta and organoids. Each pair of HLA alleles defined, at low resolution, for each sample at a particular locus is listed within the relevant column. The maternal origin of the organoids is clear from HLA-A, HLA-B and HLA-C alleles in three of the four pregnancies. The data were inconclusive for pregnancy two, but all four organoids are clearly maternal in origin when using both methods (see above).

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 3 | LC–MS/MS analysis of the secretome of trophoblast organoids

Accession	Description	-10lgP	Coverage (%)	Peptides	Unique	Spec	Placental specific	Placental enriched
Glycoproteins								
Q16557|PSG3_HUMAN	Pregnancy-specific beta-1-glycoprotein 3 GN=PSG3	272.19	28	10	1	25	Yes	
P11465|PSG2_HUMAN	Pregnancy-specific beta-1-glycoprotein 2 GN=PSG2	259.79	34	8	2	30	Yes	
Q00888|PSG4_HUMAN	Pregnancy-specific beta-1-glycoprotein 4 GN=PSG4	236.62	21	7	1	17	Yes	
P11464|PSG1_HUMAN	Pregnancy-specific beta-1-glycoprotein 1 GN=PSG1	228.57	21	8	1	15	Yes	
Classical peptides/hormones								
P01233|CGHB_HUMAN	Choriogonadotropin subunit beta GN=CGB	340.98	48	20	20	218	Yes	
P01215|GLHA_HUMAN	Glycoprotein hormones alpha chain GN=CGA	167.67	44	3	3	33	Yes	
Q15726|KISS1_HUMAN	Metastasis-suppressor KiSS-1 GN=KISS1	268.26	62	11	11	59	Yes	
Q99988|GDF15_HUMAN	Growth/differentiation factor 15 GN=GDF15	236.31	34	7	7	42		Yes
Q14641|INSL4_HUMAN	Early placenta insulin-like peptide GN=INSL4	154.72	27	2	2	4	Yes	
P35318|ADML_HUMAN	Adrenomedullin GN=ADM	114.05	10	1	1	8		Yes
P0DML2|CSH1_HUMAN	Chorionic somatomammotropin hormone 1 GN=CSH1	85.82	6	1	1	1	Yes	
P01344|IGF2_HUMAN	Insulin-like growth factor II GN=IGF2	59.42	4	1	1	1	Yes	
P01303|NPY_HUMAN	Pro-neuropeptide Y GN=NPY	50.4	8	1	1	1	unknown	unknown
Placental specific proteins (not hormones)								
P35556|FBN2_HUMAN	Fibrillin-2 GN=FBN2	157.65	1	2	2	3	Yes	
Q9BXP8|PAPP2_HUMAN	Pappalysin-2 GN=PAPPA2	124.88	1	1	1	1	Yes	
O60829|PAGE4_HUMAN	P antigen family member 4 GN=PAGE4	112.65	37	2	2	2	Yes	
P14061|DHB1_HUMAN	Estradiol 17-beta-dehydrogenase 1 GN=HSD17B1 PE=1 SV=3	83.47	5	1	1	1	Yes	
P02771|FETA_HUMAN	Alpha-fetoprotein GN=AFP PE=1 SV=1	74.38	2	2	1	19	Yes	
Q96GT9|XAGE2_HUMAN	X antigen family member 2 GN=XAGE2	62.57	23	1	1	1	Yes	
Proteins enriched in placental tissue								
P09486|SPRC_HUMAN	SPARC GN=SPARC	310.3	59	13	13	29		Yes
Q12805|FBLN3_HUMAN	EGF-containing fibulin-like extracellular matrix protein 1 GN=EFEMP1	270.65	25	7	7	17		Yes
P15121|ALDR_HUMAN	Aldose reductase GN=AKR1B1	264.8	49	8	8	21		Yes
P80723|BASP1_HUMAN	Brain acid soluble protein 1 GN=BASP1	234.76	36	4	4	13		Yes
P02751|FINC_HUMAN	Fibronectin GN=FN1	233.91	5	7	7	7		Yes
O95633|FSTL3_HUMAN	Follistatin-related protein 3 GN=FSTL3	222.26	29	5	5	27		Yes
P23142|FBLN1_HUMAN	Fibulin-1 GN=FBLN1	218.6	10	5	5	10		Yes
P10451|OSTP_HUMAN	Osteopontin GN=SPP1	178.67	32	5	5	5		Yes
P14543|NID1_HUMAN	Nidogen-1 GN=NID1	162.21	4	4	4	21		Yes
Q14118|DAG1_HUMAN	Dystroglycan GN=DAG1	148.59	5	2	2	7		Yes
P27487|DPP4_HUMAN	Dipeptidyl peptidase 4 GN=DPP4	115.06	3	1	1	1		Yes
P21980|TGM2_HUMAN	Protein-glutamine gamma-glutamyltransferase 2 GN=TGM2	106.05	6	2	2	2		Yes
P24593|IBP5_HUMAN	Insulin-like growth factor-binding protein 5 GN=IGFBP5	100.63	5	1	1	1		Yes
P09601|HMOX1_HUMAN	Heme oxygenase 1 GN=HMOX1	100.13	8	1	1	1		Yes
Q15582|BGH3_HUMAN	Transforming growth factor-beta-induced protein ig-h3 GN=TGFBI	99.46	3	1	1	1		Yes
O76061|STC2_HUMAN	Stanniocalcin-2 GN=STC2	94.72	8	1	1	2		Yes
P05111|INHA_HUMAN	Inhibin alpha chain GN=INHA	80.59	5	1	1	2		Yes
P25815|S100P_HUMAN	Protein S100-P GN=S100P	70.28	40	2	2	2		Yes
P17936|IBP3_HUMAN	Insulin-like growth factor-binding protein 3 GN=IGFBP3	58.22	6	1	1	1		Yes

Supernatants from independently derived trophoblast organoid cultures from six different placental samples were analysed: TOrg_2 (passage 23), TOrg_3 (passage 20), TOrg_5 (passage 6), TOrg_10 (passage 12), TOrg_12 (passage 4) and TOrg_14 (passage 5). The glycoproteins, classical peptides or protein hormones, placental-specific proteins and proteins enriched in placental tissue identified in the secretome data by LC–MS/MS are shown. The −10lgP score is the statistical significance assigned to a peptide or protein match by the PEAKS software[37]. 'Coverage' refers to the proportion of the primary amino acid sequence of each protein that is identified in the experiment. 'Peptides' refers to the number of peptide matches assigned to a protein; 'unique' refers to the number of peptides that are assigned solely to that protein group. 'Spec' refers to the number of peptide MS/MS spectra matched against a particular protein. The columns also indicate whether these are unique products produced by the placenta and/or whether they are products produced by other tissues but highly enriched in the placenta (among the top 10 organs that produce that protein based on RNA expression levels). Tissue location data were compared to data from the Human Protein Atlas (https://www.proteinatlas.org/). All protein identification data are in Supplementary Table 2.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0694-x

Lineage tracking reveals dynamic relationships of T cells in colorectal cancer

Lei Zhang[1,6], Xin Yu[2,6], Liangtao Zheng[1,6], Yuanyuan Zhang[3,6], Yansen Li[4], Qiao Fang[1], Ranran Gao[3], Boxi Kang[3], Qiming Zhang[3], Julie Y. Huang[2], Hiroyasu Konno[2], Xinyi Guo[3], Yingjiang Ye[4], Songyuan Gao[5], Shan Wang[4], Xueda Hu[3], Xianwen Ren[3], Zhanlong Shen[4]*, Wenjun Ouyang[2]* & Zemin Zhang[1,3]*

T cells are key elements of cancer immunotherapy[1] but certain fundamental properties, such as the development and migration of T cells within tumours, remain unknown. The enormous T cell receptor (TCR) repertoire, which is required for the recognition of foreign and self-antigens[2], could serve as lineage tags to track these T cells in tumours[3]. Here we obtained transcriptomes of 11,138 single T cells from 12 patients with colorectal cancer, and developed single T cell analysis by RNA sequencing and TCR tracking (STARTRAC) indices to quantitatively analyse the dynamic relationships among 20 identified T cell subsets with distinct functions and clonalities. Although both $CD8^+$ effector and 'exhausted' T cells exhibited high clonal expansion, they were independently connected with tumour-resident $CD8^+$ effector memory cells, implicating a TCR-based fate decision. Of the $CD4^+$ T cells, most tumour-infiltrating T regulatory (T_{reg}) cells showed clonal exclusivity, whereas certain T_{reg} cell clones were developmentally linked to several T helper (T_H) cell clones. Notably, we identified two *IFNG*$^+$ T_H1-like cell clusters in tumours that were associated with distinct IFNγ-regulating transcription factors —the *GZMK*$^+$ effector memory T cells, which were associated with *EOMES* and *RUNX3*, and *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cell clusters, which were associated with *BHLHE40*. Only *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells were preferentially enriched in patients with microsatellite-instable tumours, and this might explain their favourable responses to immune-checkpoint blockade. Furthermore, *IGFLR1* was highly expressed in both *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells and $CD8^+$ exhausted T cells and possessed co-stimulatory functions. Our integrated STARTRAC analyses provide a powerful approach to dissect the T cell properties in colorectal cancer comprehensively, and could provide insights into the dynamic relationships of T cells in other cancers.

Tumour-infiltrating lymphocytes are highly heterogeneous with respect to their cell-type compositions, gene expression profiles and functional properties, which might contribute to diverse responses to cancer immunotherapies[1]. Recent clinical trials have demonstrated that patients with colorectal cancer (CRC) who display microsatellite instability (MSI) but not microsatellite-stable (MSS) phenotypes respond to the immune-checkpoint blockade of PD-1[4], but the underlying mechanisms are not fully understood[5–7]. Several single-cell RNA sequencing (RNA-seq) studies have revealed diverse subsets and functions of T cells in various cancer types[8–10]. Here we developed an integrated approach, STARTRAC, to track further the dynamic relationships among T cell subsets identified inside colorectal carcinoma, adjacent normal mucosa and peripheral blood, based on both single-cell transcriptome and TCR α- and β-chain sequences as lineage-specific markers[2,11] (Extended Data Fig. 1a, b). STARTRAC incorporated several unique indices, including STARTRAC distribution (dist), expansion (expa), migration (migr) and transition (tran), to quantitatively describe tissue distribution, clonal expansion, migration and developmental transition or differentiation, respectively, which are essential for anti-tumour immunity by T cells (Extended Data Fig. 1b, Methods).

We obtained transcriptome data for 11,138 single T cells from 12 patients with CRC, including 4 MSI and 8 MSS patients (Extended Data Figs. 1a, 2a, Supplementary Table 1). Genomic alterations of these tumours were consistent with the characteristics of CRC from The Cancer Genome Atlas (TCGA)[12] (Extended Data Fig. 2b, Supplementary Table 2). $CD8^+$ T cells, $CD4^+CD25^{-/int}$ T_H cells and $CD4^+CD25^{hi}$ T_{reg} cells were assessed by multi-colour immunohistochemistry (IHC) (Extended Data Fig. 1c) and collected by fluorescence-activated cell sorting (FACS) before deep single-cell RNA-seq analyses (Extended Data Fig. 1a, d, Methods). Overall, we obtained an average of 1.25 million uniquely mapped read pairs (Extended Data Fig. 3a, b, Supplementary Table 3). After a series of quality control filtering, 10,805 cells remained—of which 91.4% had at least one pair of full-length productive α and β chains (Extended Data Fig. 3c, d, Supplementary Table 4). There were 7,274 clonotypes, each of which had unique productive α–β chain pairs and out of which 870 were represented by two or more cells that resulted in 3,474 clonal T cells.

A total of 8 $CD8^+$ and 12 $CD4^+$ T cell clusters were identified, each exhibiting a distinct distribution of clonotypes and clonal T cells (Fig. 1a, Extended Data Fig. 3e). The stability of clusters was supported by different clustering methods, down-sampling analysis (Extended Data Fig. 3f, g) and distinct signature genes (Extended Data Fig. 4a–c, Supplementary Table 5). In addition to typical $CD8^+$ and $CD4^+$ T cell clusters including naive (T_N), central memory (T_{CM}) and effector memory (T_{EM}) T cells, recently activated effector memory or effector T cells (T_{EMRA}/T_{EFF}, designated T_{EMRA} hereafter), mucosal-associated invariant T (MAIT) cells, blood-T_{reg} cells, tumour-T_{reg} cells, and dysfunctional or 'exhausted' $CD8^+$ T (T_{EX}) cells, we also identified two *IFNG*$^+$ T_H1-like cell clusters, with the CD4_C07-GZMK cluster expressing several markers of T_{EM} cells (designated as $CD4^+$ T_{EM} cells) and the CD4_C09-CXCL13 cluster showing higher expression of *CXCL13* and *BHLHE40* (designated as *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells; Extended Data Fig. 5a). In contrast to the hepatocellular carcinoma (HCC) and non-small-cell lung carcinoma (NSCLC) datasets[9,10], we found additional T cell subsets, including T_H17 (CD4_C08-IL23R), follicular T helper cells (CD4_C06-CXCR5), follicular T regulatory cells (CD4_C11-IL10), and two additional subsets of $CD8^+$ T cells (CD8_C05-CD6 and CD8_C06-CD160). The latter two highly expressed *CD69* and *ITGAE*, known markers of tissue-resident memory T (T_{RM}) cells[13] (Extended Data Fig. 5b). Whereas CD8_C05-CD6 probably represented lamina propria T_{RM} cells[13], CD8_C06-CD160 was characterized as intraepithelial lymphocytes (IELs) based on the highly expressed natural killer cell markers[14].

Our STARTRAC-dist index revealed distinct patterns of tissue distribution of different T cells (Fig. 1b, Extended Data Fig. 5c, d). Within the $CD8^+$ subtypes, T_N, T_{CM} and T_{EMRA} cells were predominantly enriched

[1]Beijing Advanced Innovation Centre for Genomics, Peking-Tsinghua Centre for Life Sciences, Peking University, Beijing, China. [2]Department of Inflammation and Oncology, Discovery Research, Amgen, South San Francisco, CA, USA. [3]BIOPIC and School of Life Sciences, Peking University, Beijing, China. [4]Department of Gastroenterological Surgery, Peking University People's Hospital, Beijing, China. [5]Department of Pathology, Peking University People's Hospital, Beijing, China. [6]These authors contributed equally: Lei Zhang, Xin Yu, Liangtao Zheng, Yuanyuan Zhang.
*e-mail: shenlong1977@163.com; wouyang@amgen.com; zemin@pku.edu.cn

© 2018 Springer Nature Limited. All rights reserved.

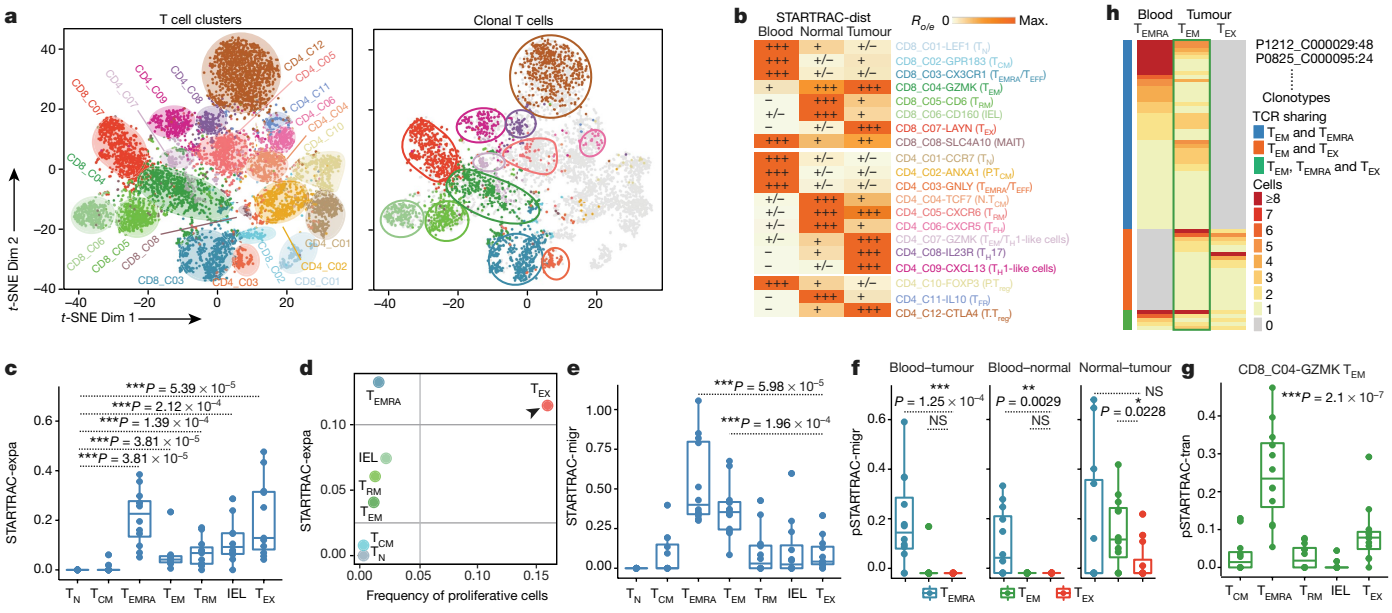


Fig. 1 | Properties of CD8$^+$ T cell clonal expansion, migration and developmental transition. a, Left, *t*-distributed stochastic neighbour embedding (*t*-SNE) plot of 8,530 T cells from 12 patients with CRC showing 20 major clusters (8 for 3,628 CD8$^+$ and 12 for 4,902 CD4$^+$ T cells; functional interpretations in Extended Data Fig. 5d). Right, highlighted clonal T cells ($n = 3{,}065$). Each dot denotes an individual T cell; colour denotes cluster origin. **b**, Tissue preference of each cluster estimated by the STARTRAC-dist index. +++, $R_{o/e} > 1$; ++, $0.8 < R_{o/e} \leq 1$; +, $0.2 \leq R_{o/e} \leq 0.8$; +/−, $0 < R_{o/e} < 0.2$; −, $R_{o/e} = 0$; in which $R_{o/e}$ denotes the ratio of observed to expected cell number. N, normal tissue; P, peripheral blood; T, tumour. **c**, Clonal expansion levels of CD8$^+$ clusters quantified by STARTRAC-expa for each patient ($n = 12$). **d**, Frequencies of proliferative CD8$^+$ T cells (*x* axis) versus the STARTRAC-expa index (*y* axis). **e**, Migration potentials of CD8$^+$ T cell clusters quantified by overall STARTRAC-migr indices for each patient ($n = 12$). **f**, Comparison of migration potentials of CD8$^+$ T_{EMRA}, T_{EM} and T_{EX} cells by pairwise STARTRAC-migr (pSTARTRAC-migr) indices for each patient ($n = 12$). **g**, Developmental transition of CD8$^+$ T_{EM} cells with other CD8$^+$ cells quantified by pairwise STARTRAC-tran indices for each patient ($n = 12$). ***$P < 0.001$, Kruskal–Wallis test. **h**, The distribution of clonal clonotypes in indicated CD8$^+$ subsets. Tumour T_{EM} cells showing mutually exclusive TCR sharing with blood T_{EMRA} and tumour T_{EX} cells (Extended Data Fig. 8e). NS, not significant. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Wilcoxon test (**c**, **e** and **f**). For box plots in all figures, centre lines denote median values; whiskers denote 1.5 × the interquartile range; coloured dots denote outliers.

in blood (Fig. 1b). T_{EX} cells were specifically enriched in tumours, whereas two subsets of T_{RM} cells were predominantly found in normal mucosa. Likewise, among CD4$^+$ subtypes, naive and effector-like cells were enriched in blood. Follicular T helper cells were enriched in normal mucosa, whereas two *IFNG*$^+$ T_H1-cell-like subsets and T_H17 cells were enriched in tumours. Three *FOXP3*$^+$ T_{reg} cell clusters, CD4_C10-FOXP3, CD4_C11-IL10 and CD4_C12-CTLA4, were enriched in blood, normal mucosa and tumours, respectively.

Focusing on CD8$^+$ T cells, the STARTRAC-expa index revealed the CD8$^+$ T_{EX} and T_{EMRA} cells as the clusters with the highest degree of clonal expansion, followed by IELs (Fig. 1c). T_{EX} cells in CRC contained the highest percentage of proliferative cells (Fig. 1d), and enriched with *MKI67*^{hi} cells and proliferation-related pathways, as confirmed by IHC (Extended Data Fig. 6a–c), although these high-proliferative cells resembled the low-proliferative cells with respect to the expression of key genes, including *TBX21*, *EOMES* and *PDCD1*[15] (Extended Data Fig. 6d, e). Consistently, although 79.26% of high-proliferative cells were clonal, most of their clonotypes (40 out of 46) were shared with low-proliferative cells (Extended Data Fig. 6f); this suggests that proliferative status was not a distinguishing feature, as was observed between progenitor and terminally differentiated exhausted states from chronic infections[15]. Furthermore, although ex vivo reactivation experiments have demonstrated that these CD8$^+$ T_{EX} cells produce less effector cytokines[16], our analyses revealed that these cells expressed higher levels of effector molecules—such as *IFNG*, *GZMB*, *GZMH* and *PRF1*—than other CD8$^+$ subsets (Extended Data Fig. 6g); this indicates that T_{EX} cells may not have completely lost their anti-tumour effector potential in vivo. Several transcription factors were preferentially expressed in the CD8$^+$ T_{EX} cell subset. Although *PRDM1* and *BATF* were the only previously known factors[17], *RBPJ*, *TOX* and *BHLHE40* were functionally uncharacterized in T_{EX} cells (Extended Data Fig. 6g).

The STARTRAC-migr analysis of CD8$^+$ clusters revealed that T_{EMRA} cells were associated with the highest mobility, followed by T_{EM} cells, and both clusters had higher mobility than T_{EX} cells (Fig. 1e, Extended Data Fig. 7a). Furthermore, pairwise STARTRAC-migr analyses revealed a high degree of TCR sharing of T_{EMRA} cells among blood, normal mucosa and tumours, whereas T_{EX} exhibited tumour exclusivity (Fig. 1f). Accordingly, these clusters expressed different sets of genes related to migration[18], including chemokine receptors, integrin and other trafficking-related molecules such as S1P receptors (Extended Data Fig. 7b). For example, T_{EMRA} cells highly expressed *S1PR1*, *S1PR5* and *ITGB7*, which supports their capability to circulate in the periphery and home to normal mucosa and tumours.

Although T_{EMRA} cells also expressed many effector molecules such as *PRF1*, *GZMB* and *GZMH*, they did not express T_{EX} cell markers such as *PDCD1* and *HAVCR2* (also known as *TIM3*) (Extended Data Fig. 6g). Notably, STARTRAC-tran analysis indicated that both T_{EMRA} and T_{EX} cell clusters were highly associated with T_{EM} cells (Fig. 1g, Extended Data Fig. 8a). T_{EX} cells were only linked to T_{EM} cells, but both T_{EMRA} and T_{EM} cells were associated with T_{CM} cells (Extended Data Fig. 8b). Furthermore, T_{EM} cells were also moderately connected to normal-enriched T_{RM} cells. Accordingly, Monocle trajectory analysis of these CD8$^+$ cell clusters also corroborated a developmental trajectory from T_{EM} cells to either T_{EMRA} or T_{EX} cells (Extended Data Fig. 8c).

The transition from T_{EM} to T_{EX} cells may predominantly occur in tumours based on their tissue distribution patterns (Fig. 1b). Although only 19.35% (24 out of 124) of T_{EMRA} cell-expanded clonotypes had clonal cells located inside tumours, 44.35% (55 out of 124) were linked to tumour-infiltrating T_{EM} cells, whereas only 5.65% (7 out of 124) were associated with blood T_{EM} cells (Extended Data Fig. 8d, Supplementary Table 6); this supports the developmental connection between T_{EMRA} cells and tumour T_{EM} cells. Notably, tumour T_{EM} cell clones linked to

© 2018 Springer Nature Limited. All rights reserved.

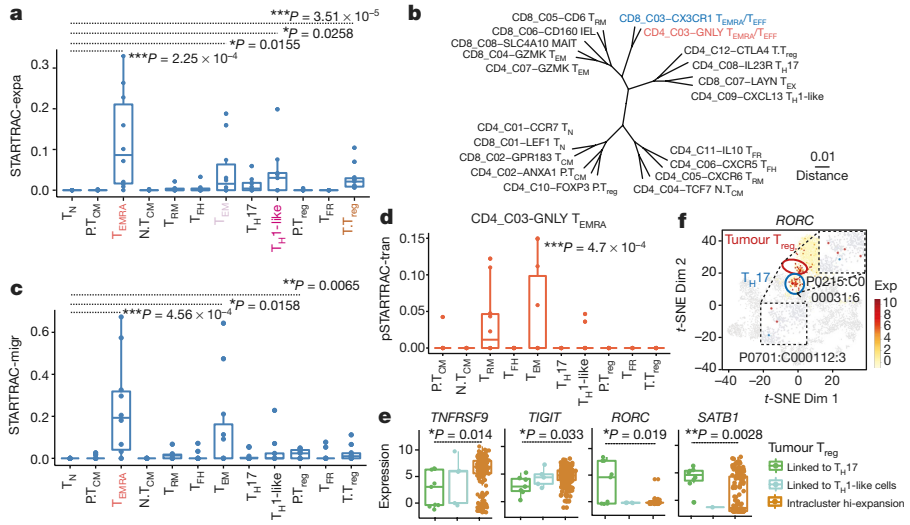


Fig. 2 | Properties of CD4$^+$ T cell clonal expansion, migration and developmental transition. a, Clonal expansion levels of CD4$^+$ T cell clusters quantified by STARTRAC-expa indices for each patient ($n = 11$). **b**, Similarities of the signature gene expressions of T cell clusters. Distance = (1 − Pearson correlation coefficient)/2. **c**, Migration potentials of CD4$^+$ T cell clusters quantified by overall STARTRAC-migr indices for each patient ($n = 11$). **d**, Developmental transition of CD4$^+$ T_{EMRA} cells with other CD4$^+$ cells quantified by pSTARTRAC-tran indices for each patient ($n = 10$). ***$P < 0.001$, Kruskal–Wallis test. **e**, Normalized expression of a series of genes in three tumour T_{reg} cell subpopulations that shared TCRs with T_H17 cells ($n = 9$ cells), *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells ($n = 5$) and exclusive to their own respective group ($n = 228$). *$P < 0.05$, **$P < 0.01$, two-sided Student's *t*-test. **f**, Representative clonotypes of tumour T_{reg} cells ($n = 1{,}320$) with high expression of *RORC* (coloured dots in ellipse) and shared TCRs with T_H17 cells ($n = 244$). Red denotes T_{reg} cells; blue denotes T_H17 cells. Exp, centred normalized expression. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Wilcoxon test (**a**, **c**).

blood T_{EMRA} cells were mutually exclusive with those linked to T_{EX} cells (Fig. 1h, Extended Data Fig. 8e). This pattern was also confirmed in individual patients (Extended Data Fig. 8f). Thus, TCR clonotypes may have a role in determining the developmental trajectories between T_{EM} and T_{EX} cells and between T_{EM} and T_{EMRA} cells.

In contrast to CD8$^+$ cells, CD4$^+$ cells exhibited lower clonal expansion overall. Among all CD4$^+$ clusters, CD4_C03-GNLY exhibited the highest clonal expansion (Fig. 2a). This cluster contained signature genes that include *NKG7*, *PRF1*, *GNLY* and *GZMH*—a signature that is profoundly similar to that of CD8$^+$ T_{EMRA} cells (Fig. 2b, Extended Data Fig. 9a); CD4_C03-GNLY cells were therefore designated as CD4$^+$ T_{EMRA} cells. Notably, the CD4_C03-GNLY cluster showed migration properties that were comparably high to those of CD8$^+$ T_{EMRA} cells (Fig. 2c), and it was linked to the CD4$^+$ T_{EM} population (Fig. 2d).

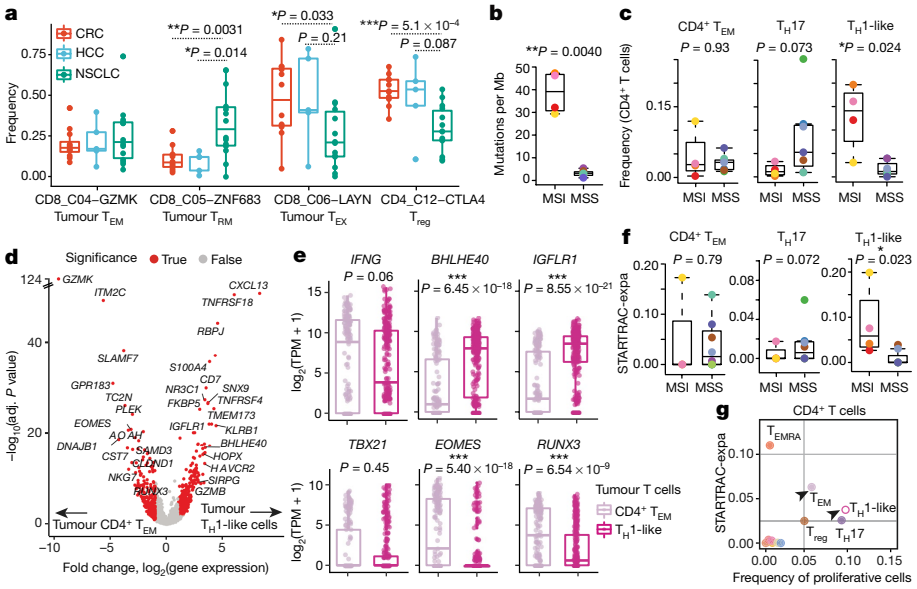


Fig. 3 | Clonal T_H1-like T cells are enriched in MSI tumours. **a**, Comparison of the proportions of different CD8$^+$ and CD4$^+$ T cell clusters in tumours from patients with CRC ($n = 12$), HCC[9] ($n = 5$), and NSCLC[10] ($n = 14$) after re-clustering of the combined dataset (Extended Data Fig. 10). Each dot denotes an individual patient. **b**, Box plot showing mutation load of patients with MSI ($n = 4$) and MSS ($n = 8$) CRC. **c**, Percentages of tumour-enriched T_H cells in the overall CD4$^+$ T cells from patients with MSI ($n = 4$) and MSS ($n = 7$) CRC. **d**, Volcano plot showing differentially expressed genes between *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells ($n = 315$) and CD4$^+$ T_{EM} cells ($n = 161$) in tumours. $P < 0.01$, Benjamini–Hochberg adjusted two-sided unpaired limma-moderated *t*-test; fold change ≥ 2. **e**, Gene expression comparison between *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells ($n = 315$) and CD4$^+$ T_{EM} cells ($n = 161$). **f**, STARTRAC-expa indices for tumour-enriched T_H cells in patients with MSI ($n = 4$) and MSS ($n = 7$) CRC. **g**, Frequencies of proliferative cells versus STARTRAC-expa index for CD4$^+$ cells. Each dot in **b**, **c** and **f** represents an individual patient. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Wilcoxon test (**a**–**c**, **e**, **f**). TPM, transcripts per million.

© 2018 Springer Nature Limited. All rights reserved.

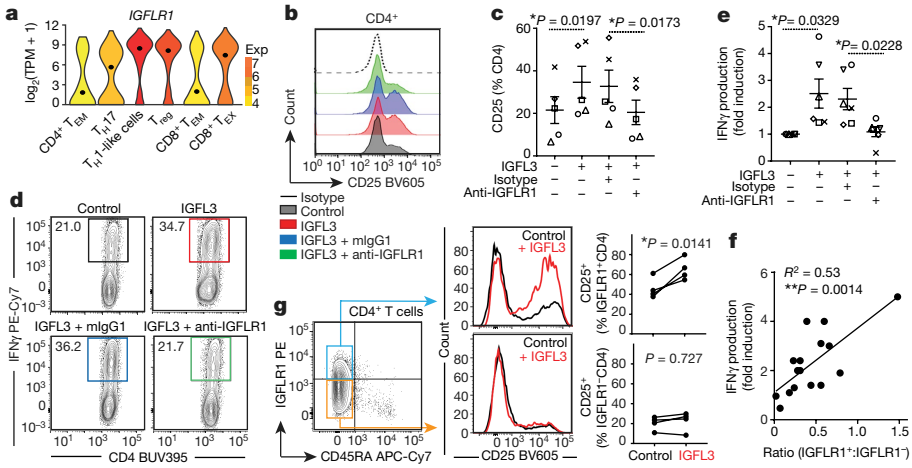


Fig. 4 | IGFLR1 functions as a co-stimulatory receptor in T cells. **a**, Violin plots showing *IGFLR1* expression in tumour-enriched T cell clusters, including CD4$^+$ T_{EM} cells ($n = 185$), T_H17 cells ($n = 244$), T_H1-like cells ($n = 319$), T_{reg} cells ($n = 1{,}320$), CD8$^+$ T_{EM} cells ($n = 773$) and CD8$^+$ T_{EX} cells ($n = 860$). Black dots denote mean values; widths denote cell densities. **b**, Representative histograms of CD25 expression in activated and indicated CD4$^+$ T cells on day 2 ($n = 5$ donors, $n = 3$ independent experiments). mIgG1 denotes mouse anti-IgG1 control antibody. **c**, Quantification of CD25 expression in **b**. Symbols are individual donors. Data are mean ± s.e.m. ($n = 5$; **c**, **e**). **d**, Flow cytometry of intracellular IFNγ expression of activated CD4$^+$ T cells on day 2 ($n = 6$ donors, $n = 3$ independent experiments). **e**, Fold changes in IFNγ supernatant from **d** measured by ELISA relative to the control group. **f**, Correlation of IFNγ induction by IGFL3 and the ratio of IGFLR1$^+$ to IGFLR1$^-$ CD4$^+$ T_{MEM} cells ($n = 16$ donors; *F*-test). **g**, Activated CD4$^+$ T_{MEM} cells were gated as IGFLR1$^+$ and IGFLR1$^-$ cells by FACS (left). Representative histograms (middle) and percentages (right) of CD25 expression levels in IGFLR1$^+$ and IGFLR1$^-$CD4$^+$ T_{MEM} cells ($n = 4$ donors). Two-sided paired Student's *t*-test (**c**, **e** and **g**).

The potential tumour-killing activities of the CD4_C03-GNLY cluster should therefore be further explored.

Tumour-infiltrating T_{reg} cells were among the highly expanded populations (Fig. 2a). Most (88%) clonal CRC-infiltrating T_{reg} cells contained TCR clonotypes exclusive to themselves, indicating their potential for recognizing tumour-associated antigens and local expansion characteristics[9,10] (Extended Data Fig. 9b). A few tumour T_{reg} cells also shared TCRs with T_{reg} cells from blood (CD4_C10-FOXP3) or normal mucosa (CD4_C11-IL10). Several other tumour T_{reg} cells shared TCRs with T_H cells in tumours (Extended Data Fig. 9b, Supplementary Table 7), indicating that they were induced T_{reg} (iT_{reg}) cells[19]. STARTRAC-tran analysis suggested that these potential iT_{reg} cells were developmentally linked to T_H17 and *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells (Extended Data Fig. 9c). iT_{reg} cells that share TCRs with T_H17 cells fell into a sub-cluster of a tumour T_{reg} cell group with higher expression of *RORC*[11,20] (Fig. 2e, f). The co-expression of RORγ and FOXP3 at the protein level was confirmed by IHC (Extended Data Fig. 9d). In addition, *SATB1*[21] was selectively expressed in T_{reg} cells linked to T_H17 cells, whereas *BACH2*[22] was preferentially expressed in T_{reg} cells linked to *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells (Fig. 2e, Supplementary Table 8). By contrast, those T_{reg} cells with high intra-cluster expansion had relatively high expression of *TNFRSF9* and *TIGIT* (Fig. 2e), suggesting that at least some of these cells might belong to natural T_{reg} cells[23]. The roles of these different subsets of T_{reg} cells needs further investigation.

When comparing T cell populations across cancer types[9,10] (Extended Data Fig. 10a, b), we found that although the composition and abundance of blood-derived T cells were highly similar, T cell patterns were distinct in both tumours and adjacent normal tissues (Extended Data Fig. 10c–f). Notably, CRC and HCC tumours exhibited a higher abundance of CD8$^+$ T_{EX} and CD4$^+$ T_{reg} cells, whereas NSCLC tumours exhibited enrichment of tumour T_{RM} cells with low expression of *PDCD1* and *CTLA4* but high expression of *ZNF683* (Fig. 3a). Likewise, the IELs were specifically present in normal mucosa and tumours in patients with CRC (Extended Data Fig. 10e, f).

Next, by focusing on the differences between heavily mutated MSI tumours (Fig. 3b) and MSS tumours, we found that MSI tumours exhibited abundant *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells, whereas MSS tumours were moderately enriched with T_H17 cells (Fig. 3c). Accordingly, the *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cell signal was increased in the TCGA MSI-high patients with CRC, whereas the T_H17 signal was enriched in the MSS cohort (Extended Data Fig. 10g). Although increased overall *IFNG*$^+$ T_H1 cells has been suggested in MSI CRC tumours[6,7], of the two *IFNG*$^+$ T_H1-like cell subsets we identified, only the *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cell cluster and not another *GZMK*$^+$*IFNG*$^+$ T_{EM} cell cluster was enriched in MSI tumours. Notably, although *TBX21* showed similar expression in both subsets, other IFNγ-regulating transcription factors *EOMES* and *RUNX3*[24] were preferentially expressed in *GZMK*$^+$ T_{EM} cells, whereas *BHLHE40* was selectively expressed in *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells (Fig. 3d, e, Supplementary Table 9); this suggests distinctive transcriptional control for these two *IFNG*$^+$ subsets. Indeed, BHLHE40 not only regulates IFNγ but also represses IL-10 production[25,26]. Furthermore, STARTRAC analyses revealed that *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells were clonally expanded and enriched in MSI patients, and proliferative in tumours (Fig. 3f, g). Moreover, these cells exhibited developmental connection with *GZMK*$^+$ T_{EM} cells, indicating the potential inter-conversion of these two *IFNG*$^+$ subsets (Extended Data Fig. 9c). Given the involvement of T_H1-like cells in response to anti-CTLA4 therapy in melanoma[27], we speculate that the enrichment of *CXCL13*$^+$*BHLHE40*$^+$*IFNG*$^+$ T_H1-like cells in MSI patients might contribute to their favourable response to immunotherapies.

These *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like cells in tumours showed high expression of *CXCR3*, *HAVCR2*, *PDCD1*, *ICOS* and *TIGIT* (Extended Data Fig. 11a, Supplementary Table 10). Notably, a less-characterized gene, *IGFLR1*, was also upregulated in these cells. IGFLR1 belongs to the TNFR superfamily[28] and has three potential ligands, with IGFL1 and IGFL3 showing high-affinity interactions. *IGFLR1* was also upregulated in tumour T_{EX} cells (Extended Data Fig. 11b, Supplementary Table 11) and T_{reg} cells (Fig. 4a). In vitro, suboptimal T cell activation was sufficient to upregulate IGFLR1 in memory CD4$^+$ T cells (Extended Data Fig. 11c–e). Importantly, IGFL3 enhanced CD25 induction and IFNγ secretion in CD4$^+$ T cells under this condition (Fig. 4b–e). The degree of IFNγ induction in total CD4$^+$ T cells was correlated with the induction of IGFLR1 in memory T cells (Fig. 4f). In addition, the IGFL3-augmented CD25 expression was observed only in IGFLR1$^+$ memory T cells, but not in the IGFLR1$^-$ cells in the same cultures (Fig. 4g). Furthermore, the synergistic effect by IGFL3 could be specifically blocked by an anti-IGFLR1 antibody (Fig. 4c–e).

© 2018 Springer Nature Limited. All rights reserved.

To study the role of IGFLR1 in exhausted $CD8^+$ T cells, we adopted a protocol for low-degree stimulation of chronic stimulation $CD8^+$ T (T_{CS}) cells (see Methods) to induce certain features of in vivo exhausted T cells (Extended Data Fig. 11f–h). These $CD8^+$ T_{CS} cells exhibited higher IGFLR1 expression than activated conventional $CD8^+$ effector T (T_{conv}) cells (Extended Data Fig. 11i, j). The addition of IGFL3 synergized the TCR-induced HAVCR2 expression[29] in these cells, which could be blocked by an anti-IGFLR1 antibody (Extended Data Fig. 11k, l). Altogether, these data suggest that IGFLR1 could synergize with TCR signalling and serve as a co-stimulatory molecule.

In summary, STARTRAC analyses unveiled various functional, migratory and developmental connections among different T cell subsets in CRC. Our data and previous findings in mouse models[30] revealed TCR-dependent trajectories for T_{EMRA} and T_{EX} cells from tumour-resident $CD8^+$ T_{EM} cells, suggesting therapeutic strategies to promote the conversion from T_{EX} to T_{EMRA} cells. The enrichment of *CXCL13*$^+$*BHLHE40*$^+$*IFNG*$^+$ T_H1-like cells in MSI patients not only provides a rationale for the high response rate to checkpoint blockade in these patients, but also solicits therapeutic focus on these cells. The compendium dataset of differentially expressed genes such as *IGFLR1*, available through an interactive portal at http://crc.cancer-pku.cn, can be used as a resource for further T cell exploration and the identification of regulatory mechanisms and therapeutic targets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0694-x.

Received: 26 February 2018; Accepted: 19 October 2018;
Published online 29 October 2018.

1. Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321–330 (2017).
2. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
3. Stubbington, M. J. T. et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
4. Le, D. T. et al. PD-1 Blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
5. Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
6. Llosa, N. J. et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discov.* **5**, 43–51 (2015).
7. Mlecnik, B. et al. Integrative analyses of colorectal cancer show immunoscore is a stronger predictor of patient survival than microsatellite instability. *Immunity* **44**, 698–711 (2016).
8. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
9. Zheng, C. et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356.e16 (2017).
10. Guo, X. et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).
11. Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).
12. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
13. Schenkel, J. M. & Masopust, D. Tissue-resident memory T cells. *Immunity* **41**, 886–897 (2014).
14. Cheroutre, H., Lambolez, F. & Mucida, D. The light and dark sides of intestinal intraepithelial lymphocytes. *Nat. Rev. Immunol.* **11**, 445–456 (2011).
15. Paley, M. A. et al. Progenitor and terminal subsets of $CD8^+$ T cells cooperate to contain chronic viral infection. *Science* **338**, 1220–1225 (2012).
16. Thommen, D. S. et al. A transcriptionally and functionally distinct $PD\text{-}1^+$ $CD8^+$ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. *Nat. Med.* **24**, 994–1004 (2018).
17. Chihara, N. et al. Induction and transcriptional regulation of the co-inhibitory gene module in T cells. *Nature* **558**, 454–459 (2018).
18. Griffith, J. W., Sokol, C. L. & Luster, A. D. Chemokines and chemokine receptors: positioning cells for host defense and immunity. *Annu. Rev. Immunol.* **32**, 659–702 (2014).
19. Adeegbe, D. O. & Nishikawa, H. Natural and induced T regulatory cells in cancer. *Front. Immunol.* **4**, 190 (2013).
20. Sefik, E. et al. Individual intestinal symbionts induce a distinct population of $ROR\gamma^+$ regulatory T cells. *Science* **349**, 993–997 (2015).
21. Beyer, M. et al. Repression of the genome organizer SATB1 in regulatory T cells is required for suppressive function and inhibition of effector differentiation. *Nat. Immunol.* **12**, 898–907 (2011).
22. Roychoudhuri, R. et al. BACH2 represses effector programs to stabilize T_{reg}-mediated immune homeostasis. *Nature* **498**, 506–510 (2013).
23. Zhang, Y. et al. Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells. *Blood* **122**, 2823–2836 (2013).
24. Djuretic, I. M. et al. Transcription factors T-bet and Runx3 cooperate to activate *Ifng* and silence *Il4* in T helper type 1 cells. *Nat. Immunol.* **8**, 145–153 (2007).
25. Yu, F. et al. The transcription factor Bhlhe40 is a switch of inflammatory versus antiinflammatory Th1 cell fate determination. *J. Exp. Med.* **215**, 1813–1821 (2018).
26. Huynh, J. P. et al. Bhlhe40 is an essential repressor of IL-10 during *Mycobacterium tuberculosis* infection. *J. Exp. Med.* **215**, 1823–1838 (2018).
27. Wei, S. C. et al. Distinct cellular mechanisms underlie anti-CTLA-4 and anti-PD-1 checkpoint blockade. *Cell* **170**, 1120–1133.e17 (2017).
28. Lobito, A. A. et al. Murine insulin growth factor-like (IGFL) and human IGFL1 proteins are induced in inflammatory skin conditions and bind to a novel tumor necrosis factor receptor family member, IGFLR1. *J. Biol. Chem.* **286**, 18969–18981 (2011).
29. Singer, M. et al. A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells. *Cell* **166**, 1500–1511.e9 (2016).
30. Schietinger, A. et al. Tumor-specific T cell dysfunction is a dynamic antigen-driven differentiation program initiated early during tumorigenesis. *Immunity* **45**, 389–401 (2016).

Acknowledgements We thank C. X. Ye for sample preparation and F. Wang, X. Zhang and J. S. Li for assistance with FACS. We thank the Computing Platform of the CLS (Peking University). This project was supported by Beijing Advanced Innovation Centre for Genomics at Peking University, Key Technologies R&D Program (2016YFC0900100), National Natural Science Foundation of China (81573022, 31530036, 91742203 and 81672375) and Amgen Corporation (USA). L.Z. was supported by the Postdoctoral Foundation of CLS.

Reviewer information *Nature* thanks N. Haining, M. Suva and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.Z., W.O. and L.Z. designed experiments. L.Z., X.Y., Q.F., R.G., Q.Z., J.Y.H., H.K. and X.G. performed the experiments. L.Z., L.T.Z., Y.Z., X.R. and X.H. analysed sequencing data. Y.L., J.Y., S.W., Y.G. and Z.S. collected clinical samples. B.K. constructed the website. L.Z., W.O. and Z.Z. wrote the manuscript with input from all authors.

Competing interests W.O., X.Y., H.K. and J.Y.H. are employees of Amgen Inc.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0694-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0694-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Z.S., W.O. or Z.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Human specimens. Twelve patients with CRC, including eight women and four men, were enrolled and pathologically diagnosed with colorectal adenocarcinoma at Peking University People's Hospital. Written informed consent was provided by all patients. This study was approved by the Research and Ethical Committee of Peking University People's Hospital and complied with all relevant ethical regulations. Fresh tumour and adjacent normal tissue samples (at least 2 cm from matched tumour tissues) were surgically resected from the above-described patients. Patients P0701, P0909, P1212, P1228, P0215, P0411, P0413, P0825, P0123 and P0309 had peripheral blood and paired tumour and adjacent normal tissues obtained, whereas patients P1012 and P1207 had only fresh tumour tissue and matched peripheral blood. Their ages ranged from 35 to 82 with a median of 67. None of them was treated with chemotherapy or radiation before tumour resection. The stages of these patients were classified according to the guidance of AJCC version 8. Among these patients, one was diagnosed at stage I, five at stage II, five at stage III, and one at stage IV. Among the four MSI-high (MSI-H) patients, three had positive lymph nodes (P0123, P0413 and P0909), and two had poorly differentiated disease (P0825 and P0909). Although we did not purposely exclude the stage IV patient, none of the MSI-H patients had distal metastasis, as evidenced by the enhanced computerized tomography (CT) results for abdomen, chest and pelvic areas before surgery. The available clinical characteristics are summarized in Supplementary Table 1. For the IGFLR1 study, human peripheral blood mononuclear cells (PBMCs) were obtained from 16 healthy donors after informed consent and authorization by the Amgen Research Blood Donor Program.

Single-cell collection. Tumours and adjacent normal tissues were cut into approximately 1-mm^3 pieces in the RPMI-1640 medium (Invitrogen) with 10% fetal bovine serum (FBS; Sciencell), and enzymatically digested with MACS Tumour Dissociation Kit (Miltenyi Biotec) for 30 min on a rotor at 37 °C, according to the manufacturer's instruction. The dissociated cells were subsequently passed through a 40-μm cell-strainer (BD) and centrifuged at 400*g* for 10 min. After the supernatant was removed, the pelleted cells were suspended in red blood cell lysis buffer (Solarbio) and incubated on ice for 2 min to lyse red blood cells. After washing twice with PBS (Invitrogen), the cell pellets were re-suspended in sorting buffer (PBS supplemented with 1% FBS).

PBMCs were isolated using HISTOPAQUE-1077 (Sigma-Aldrich) solution as previously described[9]. In brief, 3 ml of fresh peripheral blood was collected before surgery in EDTA anticoagulant tubes and subsequently layered onto HISTOPAQUE-1077. After centrifugation, lymphocyte cells remained at the plasma–HISTOPAQUE-1077 interface and were carefully transferred to a new tube and washed twice with PBS. Red blood cells were removed via the same procedure described above. These lymphocytes were re-suspended in sorting buffer.

Single-cell sorting, reverse transcription, amplification and sequencing. Single-cell suspensions were stained with antibodies against CD3, CD4, CD8 and CD25 (anti-human CD3, UCHT1; anti-human CD4, OKT4; anti-human CD8, OKT8; anti-human CD25, BC96; eBioscience) for FACS sorting, performed on a BD Aria III instrument. Single cells of different subtypes including cytotoxic T (T_C) cells, T_H cells and T_{reg} cells were enriched by gating 7AAD$^-$CD3$^+$CD8$^+$, 7AAD$^-$CD3$^+$CD4$^+$CD25$^{-/int}$ and 7AAD$^-$CD3$^+$CD4$^+$ CD25^{hi} T cells, respectively, and sorted into 96-well plates (Axygen) chilled to 4 °C, prepared with lysis buffer with 1 μl 10 mM dNTP mix (Invitrogen), 1 μl 10 μM Oligo dT primer, 1.9 μl 1% Triton X-100 (Sigma), and 0.1 μl 40 U μl^{-1} RNase Inhibitor (Takara).

The single-cell lysates were sealed and stored frozen at −80 °C immediately. Single-cell transcriptome amplifications were performed according to the Smart-Seq2 protocol[9,31]. The External RNA Controls Consortium (ERCC; Ambion; 1:4,000,000) was added into each well as the exogenous spike-in control before the reverse transcription. The amplified cDNA products were purified with 1 × Agencourt XP DNA beads (Beckman). A procedure of quality control was performed following the first round of purification, which included the detection of *CD3D* by qPCR (forward primer, 5′-TCATTGCCACTCTGCTCC-3′; reverse primer, 5′-GTTCACTTGTTCCGAGCC-3′) and fragment analysis by analyser AATI. For those single-cell samples with high quality after quality control (cycle threshold <30), the DNA products were further purified with 0.5 × Agencourt XP DNA beads, and the concentration of each sample was quantified by Qubit HsDNA kits (Invitrogen). Multiplex (384-plex) libraries were constructed and amplified using the TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme Biotech). The libraries were then purified with Agencourt XP DNA beads and pooled for quality assessment by fragment analyser. For all the 12 patients, purified libraries were analysed by an Illumina Hiseq 4000 sequencer with 150-bp pair-end reads. For patient P1207, only CD8$^+$ T cells were collected, thus this patient was not included when analysing CD4$^+$ T cells.

Primary human T cell isolation and in vitro activation. PBMCs from healthy donors were isolated by density gravity centrifugation (Ficoll-Paque PREMIUM; GE Healthcare). T cell subsets were isolated from PBMCs with appropriate magnetic beads following the manufacturer's protocol (StemCell Technologies). Isolated T cells were cultured in RPMI-1640 supplemented with 10% heat-inactivated FBS, 100 U ml^{-1} penicillin and streptomycin and 2-mercaptoethanol (all from Gibco). Human T cells were activated with anti-CD3 (UCHT1, 2 μg ml^{-1}, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml^{-1}, BD Biosciences). When indicated, 100 ng ml^{-1} recombinant human IGFL3 (Adipogen) was added in culture medium alone or with 20 μg ml^{-1} of anti-IGFLR1 blocking antibody (clone 905338, R&D Systems). The same amount of mouse IgG1 antibody (R&D Systems) was used as a control. Cells were collected and stained with indicated monoclonal antibodies (anti-human CD4, OKT4; anti-human CD8, RPA-T8; anti-human CD45RA, HI100; anti-human CCR7, G043H7; anti-human HAVCR2, F38-2E2; anti-human CD25, M-A251; anti-human IFNγ, B27; anti-human IGFLR1, 905338; Biolegend, BD Biosciences, or R&D Systems). Flow cytometry data were acquired with an LSR-II analyser and analysed using FlowJo software (Tree Star). Cytokine concentrations were measured in cell culture supernatants 40 h after stimulation with enzyme-linked immunosorbent assay (ELISA) kits specific for human IFNγ (BD Bioscience). All data were from three independent experiments with more than four donors.

In vitro CD8$^+$ T_{CS} cells. The CD8$^+$ T_{CS} cells were generated using an in vitro chronic low-degree stimulation protocol as described in previous studies[32,33]. Purified human CD8$^+$ T cells at 1 × 10^6–2 × 10^6 cells ml^{-1} were subjected to anti-CD3 (UCHT1, 0.2 μg ml^{-1}, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml^{-1}, BD Biosciences) stimulation for 3–4 days followed by at least an additional two rounds of re-stimulation every 3–4 days with anti-CD3 (UCHT1, 1 μg ml^{-1}, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml^{-1}, BD Biosciences) to generate CD8$^+$ T_{CS} cells. CD8$^+$ T_{CS} cells were then subjected to stimulation with or without human IGFL3 as described above. Cells were stained with indicated monoclonal antibodies (anti-human CD8, RPA-T8; anti-human HAVCR2, F38-2E2; anti-human PD-1, EH12.2H7; anti-human CD39, eBioA1; anti-LAG3, 305223H; anti-IGFLR1, 905338; Thermal Fisher Scientific, BD Biosciences, or R&D Systems).

T cell stimulation. Human memory CD4$^+$ T cells were isolated from healthy donor PBMCs to a purity of >94% with memory CD4$^+$ T cell isolation kit following manufacturer's protocol (Miltenyi Biotec). Isolated T cells were cultured in RPMI-1640 supplemented with 10% heat-inactivated FBS, 100 U ml^{-1} penicillin and streptomycin and 2-mercaptoethanol (all from Gibco). Human T cells were activated with anti-CD3 (UCHT1, 2 μg ml^{-1}, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml^{-1}, BD Biosciences) for 2 days. T cells were then rested in culture medium for 2 days followed by 16 h starvation in RPMI-1640 plus 0.5% FBS. Next, T cells were collected and subjected to stimulation. For TCR stimulation, T cells were incubated on ice for 30 min with 1 μg ml^{-1} anti-CD3 (OKT3, Thermo Fisher Scientific) and 1 μg ml^{-1} anti-CD28 followed by a 15-min incubation with 5 μg ml^{-1} anti-mouse IgG (Thermo Fisher Scientific). Cells were activated by incubation in a 37 °C water bath for 25 min. For IGFL3 stimulation, T cells were incubated on ice for 30 min with 100 ng ml^{-1} recombinant human IGFL3 (Adipogen). Cells were then activated by incubation in a 37 °C water bath for 25 min.

Cytokine production detection. Human memory CD4$^+$ T cells were isolated as described above. Cells were activated with anti-CD3 (UCHT1, 2 μg ml^{-1}, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml^{-1}, BD Biosciences) or anti-CD3 and anti-CD28 plus IGFL3 (100 ng ml^{-1}). Cells were seeded in flat-bottom 96-well plates with 10^5 cells in 100 μl culture medium per 96 wells. Cytokine concentrations were measured in cell culture supernatants 40 h after stimulation with ELISA kits specific for human IFNγ (BD Bioscience).

Bulk DNA and RNA isolation and sequencing. Genomic DNA of peripheral blood and tissue samples of patients with CRC were extracted using the QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's specification. The concentrations of DNA were quantified using the Qubit HsDNA Kits (Invitrogen) and the qualities of DNA were evaluated with agarose gel electrophoresis. Exon libraries were constructed using the SureSelectXT Human All Exon V5 capture library (Agilent). Samples were sequenced on the Illumina Hiseq 4000 sequencer with 150-bp paired-end reads. For bulk RNA analysis, small fragments of tumour tissues and adjacent normal tissues were first stored in RNAlater RNA stabilization reagent (QIAGEN) after surgical resection and kept on ice to avoid RNA degradation. RNA of tumour and adjacent normal tissue samples were extracted using the RNeasy Mini Kit (QIAGEN) according to the manufacturer's specification. The concentrations of RNA were quantified using the NanoDrop instrument (Thermo) and the qualities of RNA were evaluated with fragment analyser (AATI). Libraries were constructed using NEBNext Poly (A) mRNA Magnetic Isolation Module kit (NEB) and NEBNext Ultra RNA Library Prep Kit for Illumina Paired-end Multiplexed Sequencing Library (NEB). Samples were sequenced on the Illumina Hiseq 4000 sequencer with 150-bp paired-end reads.

© 2018 Springer Nature Limited. All rights reserved.

Microsatellite instability testing. DNA purified from tumour tissues using QIAamp DNA Mini Kit (QIAGEN) was subjected to multiplex fluorescent PCR-based assay (Promega) by amplifying seven loci including five mononucleotide repeats (NR21, BAT26, BAT25, NR24 and Mono27) and two pentanucleotide repeats (PentaC and PentaD) and was compared with DNA extracted from matched adjacent normal tissues. Multiplex PCR products were analysed by ABI PRISM 3100 Genetic Analyzer (Applied Biosystems). Patients were defined as MSI-H status by the presence of two or more mononucleotide loci showing instability. MSS was defined as no loci of instability. Among 12 patients in this study, 4 of them were MSI-H (P0413, P0825, P0123 and P0909), and the other 8 were MSS (P0215, P0411, P0701, P1012, P1207, P1212, P1228 and P0309).

Immunohistochemistry. The specimens were collected from Peking University People's Hospital within 30 min of the tumour resection and fixed in 10% formalin for 48 h. Dehydration and embedding in paraffin was performed as the following routine methods[9]. These paraffin blocks were cut into 5-μm sections and adhered to a glass slide. Then, the paraffin sections were placed in the 70-°C paraffin oven for 1 h before being deparaffinised in xylene and then rehydrated in 100%, 90% and 70% alcohol successively. The antigens were retrieved by the Epitope Retrieval Solution (Leica Biosystems), and the sections were incubated with ready-to-use primary antibodies (mouse anti-human MLH1, clone ES05; mouse anti-human MSH2, clone 25D12; mouse anti-human MSH6, clone PU29; mouse anti-human PMS2, clone M0R4G, all from Leica Biosystems) on the BOND system (Leica Biosystems) according to the manufacturer's protocol.

Multi-colour immunohistochemistry. The specimens were collected and prepared for the formalin-fixed paraffin-embedded tissues sections as previously mentioned[9]. The confirmation of RORγ$^+$ T_{reg} cells was analysed using Opal 7-Colour Manual IHC Kit (PerkinElmer, NEL811001KT) according to the manufacturer's protocol, as previously described[10]. In brief, antigen was retrieved by AR9 buffer (pH 6.0, PerkinElmer) and boiled in the oven for 15 min. After a pre-incubation with blocking buffer at room temperature for 10 min, the sections were incubated at room temperature for 1 h with rabbit anti-human CD3 (Abcam, clone SP7, 1:100), rabbit anti-human RORγ (Abcam, 1:50), and mouse anti-human FOXP3 (Abcam, clone mAbcam22510, 1:100). A secondary horseradish peroxidase-conjugated antibody (PerkinElmer) were added and incubated at room temperature for 10 min. Signal amplification was performed using TSA working solution diluted at 1:100 in 1× amplification diluent (PerkinElmer) and incubated at room temperature for 10 min. The other validations by multi-colour IHC were performed using the same protocols with different primary antibodies as follows. The primary antibodies and IHC metrics used in the validation of T_C, T_H and T_{reg} cells were rabbit anti-human CD3 (Abcam, clone SP7, 1:400), rabbit anti-human CD4 (Abcam, clone EPR6855, 1:400), mouse anti-human CD8 (Abcam, clone 144B, 1:500) and mouse anti-human FOXP3 (Abcam, clone mAbcam22510, 1:500). The primary antibodies and IHC metrics used in the validation of proliferative CD8$^+$ T_{EX} cells were: rabbit anti-human TIM-3 (also known as HAVCR2) (Cell Signaling, clone D5D5R, 1:100), mouse anti-human Ki67 (Abcam, B126.1, 1:200), mouse anti-human PD-1 (Abcam, clone NAT105, 1:200) and mouse anti-CD8 (Abcam, clone 144B, 1:200). The multispectral imaging was collected by Mantra Quantitative Pathology Workstation (PerkinElmer, CLS140089) at 20× magnification and analysed by InForm Advanced Image Analysis Software (PerkinElmer) version 2.3. For each patient, a total of 8–15 high-power fields were taken based on their tumour sizes.

Quality control and preprocessing of single-cell RNA-seq data. Low-quality read pairs of single-cell RNA sequencing (scRNA-seq) data were filtered out if at least one end of the read pair met one of the following criteria: (1) '*N*' bases account for ≥10% of the read length; (2) bases with quality <5 account for ≥50% of the read length; and (3) the read contains adaptor sequence. The filtered read pairs were processed using HTSeqGenie pipeline (R package version 4.8) to obtain the gene expression table. Specially, read pairs were then mapped to human ribosomal RNA sequences (download from RFam database) and the read pairs with both ends unmapped were kept for downstream analysis. Read pairs passing this filter for rRNA were aligned to human reference sequence (hg19) using GSNAP[34], with parameters '–novelsplicing 1 -n 10 -i 1 -M 2'. To calculate the expression levels of genes, the gene model file 'knownGene.txt' (30 June 2013 version), downloaded from UCSC, was used. The R function findOverlaps was used to count the number of uniquely mapped read pairs located in each gene and the count table tabulated as genes by cells was used for downstream analysis. The transcripts per million (TPM) table was derived from the count table and the TPM value was calculated by

$$\frac{10^6 \times C_{ij}/\text{length of gene } i}{\sum_i C_{ij}/\text{length of gene } i}$$

in which C_{ij} was the count value of gene i in cell j.

Low-quality cells were filtered if the library size or the number of expressed genes (counts larger than 0) was smaller than predefined thresholds. Both thresholds were defined as the medians of all cells minus 3 × the median absolute deviation. Furthermore, if the proportion of mitochondrial gene counts was larger than 10%, these cells were discarded. Only cells with the average TPM of *CD3D*, *CD3E* and *CD3G* larger than 10 were kept for subsequent analysis. We further identified CD4$^+$, CD8$^+$, CD4$^-$CD8$^-$ (double negative) and CD4$^+$CD8$^+$ (double positive) T cells based on the gene expression data. Given the average TPM of *CD8A* and *CD8B*, one cell was considered as CD8 positive or negative if the value was larger than 30 or less than 3, respectively; given the TPM of *CD4*, one cell was considered as CD4 positive or negative if the value was larger than 30 or less than 3, respectively. Hence, the cells can be in silico classified as CD4$^+$CD8$^-$, CD4$^-$CD8$^+$, CD4$^+$CD8$^+$, CD4$^-$CD8$^-$ and other cells that cannot be clearly defined. A total of 52 cells were filtered out owing to the inconsistent classifications based on transcriptome data and FACS.

After discarding genes with average counts of fewer than or equal to 1, the count table of the cells passing the above filtering was normalized using a pooling strategy implemented in the R function computeSumFactors[35]. With this strategy, size factors for individual cells were deconvoluted from size factors of pools, the sizes of which ranged from 20, 40, 60, 80 to 100. To avoid violating the assumption that most genes were not differentially expressed, hierarchical clustering based on Spearman's rank correlation was performed first, then normalization was performed in each cluster separately. The size factor of each cluster was further re-scaled to enable comparison between clusters. The normalized data were in $\log_2$ space. To remove the possible effects of different donors on expression, the normalized table was further centred by patient. Thus, in the centred expression table, the mean values of the cells for each patient were zero. A total of 12,548 genes and 10,805 cells were retained in the final expression table. If not explicitly stated, 'normalized read count' or 'normalized expression' in this study refers to the normalized and centred count data for simplicity.

Analysis pipelines of bulk exome sequencing and RNA-seq data. The bulk exome sequencing data were cleaned following the same procedure for the scRNA-seq data processing. The cleaned read pairs were then processed according to the BWA-PICARD/GATK-strelka pipeline. In brief, the cleaned read pairs were aligned to human genome reference version b37 (downloaded from ftp://ftp.broadinstitute.org:/bundle) by the BWA-MEM algorithm[36]. The alignments were then sorted and de-duplicated by PICARD (Broad Institute). GATK[37] was used to realign multiple reads around putative INDEL by Smith–Waterman alignment algorithm and re-calibrate base quality. The analysis-ready bam files were input into the GATK UnifiedGenotyper module to call SNP/INDEL and into strelka[38] to call somatic SNV/INDEL and into ADTEx[39] (version 1.0.4) to call somatic copy number alterations. The mutations were annotated with annovar[40].

TCR assembly. TraCeR was used to deduce the TCR sequences of each cell[3]. The outputs of TraCeR include the assembled nucleotide sequences for both α and β chains, the coding potential of the nucleotide sequences (that is, productive or not), the translated amino acid sequence, the CDR3 sequences and the estimated TPM value of α or β chains. Only cells with TPM values larger than 10 for the α chain and larger than 15 for the β chain were kept.

For cells with two or more α or β chains assembled, the α–β pair that was productive and of the highest expression level was defined as the dominant α–β pair in the corresponding cell. If two cells had identical dominant α–β pairs, the dominant α–β pair were identified as clonal TCRs. To integrate with the gene expression data, the TCR-based analysis was performed only for cells that passed the aforementioned quality control pipeline (total 10,805). Thus, 9,878 cells with TCR information were used in the integrative analysis (Supplementary Table 4).

If one cell had an α chain composed of V segment TRAV1-2 and one of the following J segments (TRAJ33, TRAJ20 and TRAJ12), the cell was classified as a MAIT cell[41]. If the α chain of one cell was rearranged by V segment TRAV10 and J segment TRAJ18, the cell was classified as an invariant natural killer T cell[42]. In the 9,878 cells with at least one pair of productive α and β chains, only 3 cells were identified as invariant natural killer T cells, and 102 cells were identified as MAIT cells, including 71 CD8$^+$CD4$^-$ T cells classified in silico.

Unsupervised clustering analysis of CRC scRNA-seq dataset. The expression tables of CD8$^+$CD4$^-$ T cells and CD8$^-$CD4$^+$ T cells as defined by the aforementioned in silico classification but excluding MAIT cells and invariant natural killer T cells, were fed into an iteratively unsupervised clustering pipeline separately. Specifically, given an expression table, the top n genes with the largest variance were selected, and then the expression data of the n genes were analysed by single-cell consensus clustering (SC3)[43]. n was tested from 500, 1,000, 1,500, 2,000, 2,500 and 3,000. In SC3, the distance matrices were calculated based on Spearman correlation and then transformed by calculating the eigenvectors of the graph Laplacian. Then, the k-means algorithm was applied to the first d eigenvectors multiple times, in which d was chosen as between 4% and 7% of the total number of input cells. Finally, hierarchical clustering with complete agglomeration was performed on the SC3 consensus matrix and k clusters were inferred. The SC3 parameter k, which was used in the k-means and hierarchical clustering, was tried from 2 to 10.

© 2018 Springer Nature Limited. All rights reserved.

For each SC3 run, the silhouette values were calculated, the consensus matrix was plotted and cluster-specific genes were identified. Such information was used to determine the optimal k and n. Once the stable clusters were determined, the above procedure was iteratively applied to each of these clusters to reveal the sub-clusters. The in silico classified $CD8^+CD4^-$ MAIT cells had distinct gene expression patterns compared with other $CD8^+CD4^-$ T cells, and were defined as cluster 'CD8_C08-SLC4A10'.

When the clustering results were obtained, one-way ANOVA implemented by R function 'aov' was performed to identify the differentially expressed genes among the clusters. R function TukeyHSD was used to identify which cluster pairs showed a significant difference. A gene was defined as being significantly differentially expressed based on the following criteria: (1) adjusted P value (Benjamini–Hochberg method) of F-test of less than 0.05; (2) the absolute difference of any one significant cluster pair (P value of Tukey's honest significant difference method less than 0.01) larger than 1. The significantly differentially expressed genes were categorized in the cluster that showed the highest expression (Supplementary Table 5).

The t-SNE method implemented in R package Rtsne was used for clustering visualization. To visualize the cell density on the t-SNE plot, kernel density estimation was performed using R function 'kde' (ks package), and the contour lines encompassing the top 10%, 20%, …90% cells with highest densities were shown. A total of 8,530 T cells, including 3,628 $CD8^+CD4^-$ and 4,902 $CD8^-CD4^+$ T cells with clustering definitions, were used in the t-SNE projection. Other cells such as $CD8^+CD4^+$ and $CD8^-CD4^-$ T cells were not included in this visualization.

To validate the clustering results from the SC3 pipeline, we also performed clustering analysis using two additional pipelines, Seurat[44] and sscClust[10]. Raw read-count tables were provided to the Seurat pipeline. For each cell, the counts were normalized by the total counts then multiplied by a scale factor of 100,000 before transforming to the $\log_2$ scale. To identify highly variable genes, the relationship between mean expression and dispersion was fitted with log(VMR) (variance to mean ratio) as dispersion function. The mean expression cut-offs were set at 0.0125 and 8 for low and high limit, respectively, and dispersion cut-off was set at 0.5 for low limit. The donor covariate effect was removed by regression, and the resulting data were used to perform PCA. The top 15 principal components were kept and clusters were identified by the SNN algorithm. Resolution parameters of 0.7 and 1.0 were set for $CD8^+$ T cell data and $CD4^+$ T cell data, respectively.

The sscClust method is a two-round clustering pipeline that uses both PCA and t-SNE for dimension reduction and uses a density-based clustering method. The normalized expression data used in SC3 analysis were also used in the sscClust analysis. In the first round, the top 1,500 genes with the highest standard deviation were used for PCA, and top principal components were used for t-SNE, implemented in R package Rtsne. The R package densityClust[45] was used for density peak identification and cluster assignment. Subsequently, differentially expressed genes were identified by analysis of variance, and the top 1,000 genes were used in the second round of PCA/t-SNE/densityClust analysis. In both rounds, top 15 principal components were used. The rho and delta parameters of densityClust, which denote the density of each cell and the minimum distance to other cells with density larger than the cell in consideration, were chosen based on the rho-delta decision plot. Specifically, for the $CD8^+$ T cell data, rho/delta of 40/5 and 30/3 were used for the first and second round analysis, respectively; for the $CD4^+$ T cell data, 50/4 and 50/4 were used for first and second round analysis, respectively.

Down-sampling analysis of CRC T cell scRNA-seq dataset. To evaluate the effect of cell numbers on clustering results, we iteratively repeated the clustering analysis after down-sampling the CRC T cell data to 100, 200, 500, 1,000, 1,500, 2,000 and 3,000 cells. For each down-sampling number, 10 replicates were performed. Each down-sampled dataset was used for clustering analysis by sscClust (for speed considerations and similarity to the SC3 results), the resulting cluster labels were compared with our benchmark labels, as obtained from the whole dataset analysis, using the normalized mutual information (NMI) index[46]. A higher NMI index means more accurate cluster assignment in the down-sampled dataset. The sscClust pipeline was run with largely the same procedure described in the previous section but with different (rho and delta) parameters. The same (rho and delta) parameters were used in both rounds of clustering. The rho parameter was set at 1.5, 3, 7.5, 15, 20, 20 and 20 for cell numbers ranging from 100 to 3,000, respectively. For the delta parameter, values of 2 to 7 were tested and the value that gave the highest NMI value was chosen as the optimal parameter.

Analysis of combined CRC, HCC and NSCLC T cell scRNA-seq datasets. The expression data of cells passing the quality control filters in the three studies for HCC, NSCLC and CRC T cells (GEO accessions GSE98638, GSE99254 and GSE108989) were fetched, and then re-processed using the same aforementioned pipeline for in silico classification and normalization. The in silico-classified $CD8^+CD4^-$ and $CD8^-CD4^+$ T cells, excluding invariant natural killer T cells, were used in combined clustering analysis. The large combined dataset demanded the use of the sscClust method for its high computational efficiency. For $CD8^+$ T cell data, rho and delta parameters (100, 5) and (100, 3) were used for the first and second round analysis respectively, and for $CD4^+$ T cell data, (120, 2) and (130, 2) were used for the first and second round analysis, respectively.

Gene set enrichment analysis. Pre-ranked analysis module in GSEA[47] was used for gene set enrichment analysis. The gene sets we used were from the database MSigDB[48]. In single cluster enrichment analysis, the normalized and centred expression data were transformed to z-scores. For each cluster, the z-scores across cells were averaged per gene. Each cluster had an average expression profile. The z-score profile was used as input for GSEA.

Identification of proliferative cells. The average expression of known proliferation-related genes was defined as the proliferation score. These proliferation genes include *ZWINT*, *E2F1*, *FEN1*, *FOXM1*, *H2AFZ*, *HMGB2*, *MCM2*, *MCM3*, *MCM4*, *MCM5*, *MCM6*, *MKI67*, *MYBL2*, *PCNA*, *PLK1*, *CCND1*, *AURKA*, *BUB1*, *TOP2A*, *TYMS*, *DEK*, *CCNB1* and *CCNE1*[49]. Proliferative cells were identified using an outlier detection procedure implemented in the R package extremevalues. Specifically, as most of the proliferation scores came from low-proliferation cells, a normal distribution was fitted using cells with proliferation scores between 10% and 90% quantiles. Cells with a proliferation score larger than a threshold were classified as proliferative cells, and this threshold value was optimally set by the getOutliersI function of the extremevalues package.

Trajectory analysis. To characterize the potential process of T cell functional changes and determine the potential lineage differentiation between diverse T cells, we applied the Monocle (version 2) algorithm[50] with the top 700 signature genes of $CD8^+$ T cells excluding MAIT cells, based on the rank of F statistic generated by ANOVA (Supplementary Table 5). Cells were ordered through the inferred pseudotime to indicate their differentiation progress. The Monocle function relative2abs was used to convert TPM measurement into mRNAs per cell (RPC), and then the CellData Set object was created with the parameter 'expressionFamily = negbinomial'. Then the $CD8^+$ T cell differentiation trajectory was inferred after dimension reduction and cell ordering with the default parameters of R package Monocle.

TCGA data analysis. The TCGA colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) data were used to confirm the differences of the T cell subtype compositions between patients with MSI-H ($n = 62$) and MSS ($n = 286$). None of the patients from the TCGA COAD and READ had any previous record of immunotherapy treatment. The gene expression data and clinical data were downloaded from UCSC Xena (http://xena.ucsc.edu/). We calculated the average expression of known marker genes of T_H17 cells (*IL17A*, *IL17F*, *IL23R*, *CCR6*, *RORC* and *CD4*) and T_H1-like cells (*CXCL13*, *HAVCR2*, *IFNG*, *CXCR3*, *BHLHE40* and *CD4*) after z-score normalization with log-transformed expression profiles. P values from the Wilcoxon test were used to determine the statistical significance in R.

Definition of STARTRAC indices for tissue distribution, clonal expansion, tissue migration and state transition. We present STRATRAC as a framework, defined by four indices, to analyse different aspects of T cells based on paired single-cell transcriptomes and TCR sequences. The first index, STARTRAC-dist, uses the ratio of observed over expected cell numbers in tissues to measure the enrichment of T cell clusters across different tissues. Given a contingency table of T cell clusters by tissues, we first apply chi-squared test to evaluate whether the distribution of T cell clusters across tissues significantly deviates from random expectations. We then calculate the STARTRAC-dist index for each combination of T cell clusters and tissues according the following formula:

$$I_{\text{dist}}^{\text{STARTRAC}} = R_{o/e} = \frac{\text{observed}}{\text{expected}}$$

in which $R_{o/e}$ is the ratio of observed cell number over the expected cell number of a given combination of T cell cluster and tissue. The expected cell number for each combination of T cell clusters and tissues are obtained from the chi-squared test. Different from the chi-squared values, which are defined as $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ and can only indicate the divergence of observations from random expectations, $I_{\text{dist}}^{\text{STARTRAC}}$ defined by $R_{o/e}$ can indicate whether cells of a certain T cell cluster are enriched or depleted in a specific tissue. For example, if $R_{o/e} > 1$, it suggests that cells of the given T cell cluster are more frequently observed than random expectations in the specific tissue, that is, enriched. If $R_{o/e} < 1$, it suggests that cells of the given T cell cluster are observed with less frequency than random expectations in the specific tissue, that is, depleted. By calculating the STARTRAC-dist indices via $R_{o/e}$, we can quantify the tissue preference of T cell clusters efficiently.

The other three STARTRAC indices, STARTRAC-expa, STARTRAC-migr and STARTRAC-tran, are designed to measure the degree of clonal expansion, tissue migration, and state transition of T cell clusters upon TCR tracking, respectively. The MAIT cells (CD8_C08-SLC4A10) were not included in these types of analyses because they have distinct TCRs. For STARTRAC-expa, which uses the standard TCR clonality measurement[51] but is specifically applied to different T cell clusters in our analyses, we first adopt the normalized Shannon entropy to calculate the evenness of the TCR repertoire of the given T cell cluster and then define the

© 2018 Springer Nature Limited. All rights reserved.

STARTRAC-expa index as 1 − evenness. Mathematically, the STARTRAC-expa index of a specific cluster with N clonotypes is defined by the following formula:

$$I_{\text{expa}}^{\text{STARTRAC}} = 1 - \text{evenness} = 1 - \frac{-\sum_{i=1}^{N} p_i \log_2 p_i}{\log_2 N}$$

in which p_i is the cell frequency of clonotype i in the cluster, and a clonotype is defined by identical, full-length, paired α and β TCR chains. Although the definition of STARTRAC-expa is mathematically identical to the clonality scores frequently used in bulk TCR repertoire sequencing studies[51], two distinctions should be noted. First, STARTRAC-expa is defined for T cell clusters while the traditional TCR clonality is defined for specific specimens. A T cell cluster in STARTRAC framework can consist of T cells from several tissues and patients, but the specimens subject to bulk TCR repertoire sequencing are typically from a unique tissue and patient. Second, STARTRAC-expa uses a more stringent clonotype definition. For traditional bulk TCR sequencing studies, clonotypes are generally defined based on identical CDR3 (the complementarity determining region 3) sequences of TCR α or β chains, owing to technological limitations. However, STARTRAC-expa is defined using the strictest clonotype definition, which requires that both the full-length α and β chains of TCRs are identical at the nucleotide level. Thus, although STARTRAC-expa has an identical mathematical formula to that of the traditional TCR clonality definition, they have distinct biological meanings. STARTRAC-expa ranges from 0 to 1, with 0 indicating no clonal expansion for each clonotype while 1 indicating that the cluster is composed of only one clonally expanded clonotype. If a cluster is composed of multiple clonotypes and each clonotype is subject to distinct extent of clonal expansion, STARTRAC-expa will be between 0 and 1, with high STARTRAC-expa indicating high clonality.

Even if T cells with identical TCR clonotypes are present in different tissues or in different development states, logically they could likely derive from a single naive T cell, clonally expanded initially at one location and migrated across tissues, or have undergone state transitions. Based on this principle, we define STARTRAC-migr and STARTRAC-tran to evaluate the extent of tissue migration and state transition of each clonotype, respectively. For each clonotype, given its distribution across tissues (peripheral blood, adjacent normal mucosa and tumour), we define its STARTRAC-migr index I_{migr}^{t} as:

$$I_{\text{migr}}^{t} = -\sum_{j=1}^{J} p_j^t \log_2 p_j^t$$

in which p_j^t is the ratio of the number of cells with TCR clonotype t in tissue j to the total number of cells with TCR clonotype t and $\sum_{j=1}^{J} p_j^t = 1$. For two T cell clusters with similar clonal expansion and clonal size, the one with clonal cells broadly distributed in various tissues would probably be more mobile. Similarly, the STARTRAC-tran index I_{tran}^{t} can be defined as:

$$I_{\text{tran}}^{t} = -\sum_{k=1}^{K} p_k^t \log_2 p_k^t$$

in which p_k^t is the ratio of the number of cells with TCR clonotype t in cluster k to the total number of cells with TCR clonotype t, $\sum_{k=1}^{K} p_k^t = 1$, and K is the total number of cell clusters. Although both definitions use Shannon entropy for calculation, they are distinct from the measurement of TCR clonality in bulk TCR repertoire sequencing. As described above, the traditional TCR clonality is defined at the sample level; however, STARTRAC-migr and STARTRAC-tran are defined primarily at the clonotype level. Given one clonotype, the evenness or diversity of its TCR repertoire will be zero because all the cells have identical TCRs, while the STARTRAC-migr and STARTRAC-tran indices will be non-trivial because cells of the same clonotype can migrate across tissues or change their transcriptional states. Thus, the inputs of the formulas of STARTRAC-migr and STARTRAC-tran are also different from the traditional TCR clonality measurement and STARTRAC-expa. The input of STARTRAC-migr is the observed cell frequency across tissues of a certain clonotype, while the input of STARTRAC-tran is the observed cell frequency across cell clusters of a certain clonotype. By contrast, the input of STARTRAC-expa is the observed cell frequency across clonotypes of a certain cell cluster, and the input for the traditional TCR clonality measure is the observed sequence frequency across a TCR repertoire of a given sample. For the calculation of STARTRAC-migr, to exclude the possible influence of different extent of expansion or local proliferation of T cells, we also calculate the proliferation-normalized STARTRAC-migr index, which normalizes the number of expanded cells of clonotypes in each tissue as 1. As expected, we found a similar trend of T cell migration potentials for both $CD8^+$ and $CD4^+$ T cells evaluated by this proliferation-normalized STARTRAC-migr as those calculated by STARTRAC-migr (data not shown). To make our calculation consistent, we used STARTRAC-migr for subsequent analyses.

After the extent of tissue migration of each clonotype is quantified by STARTRAC-migr, given a cluster with total T clonotypes, the STARTRAC-migr index at the cluster level $I_{\text{migr}}^{\text{STARTRAC}}$ can be defined as the weighted average of all TCR clonotype migration indices contained in the cluster:

$$I_{\text{migr}}^{\text{STARTRAC}} = \sum_{t=1}^{T} p_{\text{cls}}^{t} I_{\text{migr}}^{t}$$

in which p_{cls}^{t} is the ratio of the number of cells with clonotype t in cluster cls to the total number of cells in cluster cls.

Similarly, when the extent of state transition of each clonotype is quantified by STARTRAC-tran, given a cluster with total T clonotypes, the STARTRAC-tran index at the cluster level can be defined as the weighted average of all TCR clonotypes state transition indices contained in the cluster:

$$I_{\text{tran}}^{\text{STARTRAC}} = \sum_{t=1}^{T} p_{\text{cls}}^{t} I_{\text{tran}}^{t}$$

in which p_{cls}^{t} is the ratio of the number of cells with clonotype t in cluster cls to the total number of cells in cluster cls.

Of note, both STARTRAC-migr and STARTRAC-tran are defined at two different levels (clonotypes and clusters), with the clonotype-level definitions describing the extent of migration and state transition of a given clonotype, and the cluster-level definitions depicting the summarization of such properties of all clonotypes within a cluster.

Besides the overall evaluation of the extents of migration and state transitions by STARTRAC-migr and STARTRAC-tran, we also define pairwise STARTRAC-migr (pSTARTRAC-migr) and STARTRAC-tran (pSTARTRAC-tran) indices for precise quantification. For example, given a clonotype t and two tissue types (for example, blood and tumour), the pSTARTRAC-migr index ${}_{p}I_{\text{migr}}^{t}$ is calculated by the following formula:

$${}_{p}I_{\text{migr}}^{t} = -\sum_{j=1}^{2} p_j^t \log_2 p_j^t$$

in which p_j^t is the ratio of the number of cells with TCR clonotype t in tissue j to the total number of cells with TCR clonotype t in tissues 1 and 2 (that is, blood and tumour), and $\sum_{j=1}^{2} p_j^t = 1$. In other words, pSTARTRAC-migr uses the same formula as STARTRAC-migr but limits the number of tissues to two and the frequencies of cells between two specified tissues are re-calculated. Likewise, given a clonotype t and two T cell clusters (for example, T_{EM} and T_{EX} cells), the pSTARTRAC-tran index ${}_{p}I_{tran}^{t}$ is calculated by the following formula:

$${}_{p}I_{\text{tran}}^{t} = -\sum_{k=1}^{2} p_k^t \log_2 p_k^t$$

in which p_k^t is the ratio of the number of cells with TCR clonotype t in cluster k to the total number of cells with TCR clonotype t in clusters 1 and 2 (that is, T_{EM} and T_{EX} cells), and $\sum_{k=1}^{2} p_k^t = 1$. Thus, pSTARTRAC-tran uses the same formula as STARTRAC-tran but limits the number of clusters to two, and the frequencies of cells between the two specified clusters are re-calculated. Once pairwise STARTRAC-migr and STARTRAC-tran for clonotypes are obtained, the corresponding indices for clusters are calculated via weighted average according to their clonotype compositions. As all STARTRAC-migr and STARTRAC-tran indices are defined by Shannon entropy, high values indicate high migration and state transition, respectively.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The open source code is available at GitHub. Code for sscClust clustering is available on GitHub (https://github.com/Japrin/sscClust). Code for STARTRAC analysis is available on GitHub (https://github.com/Japrin/STARTRAC).

Data availability

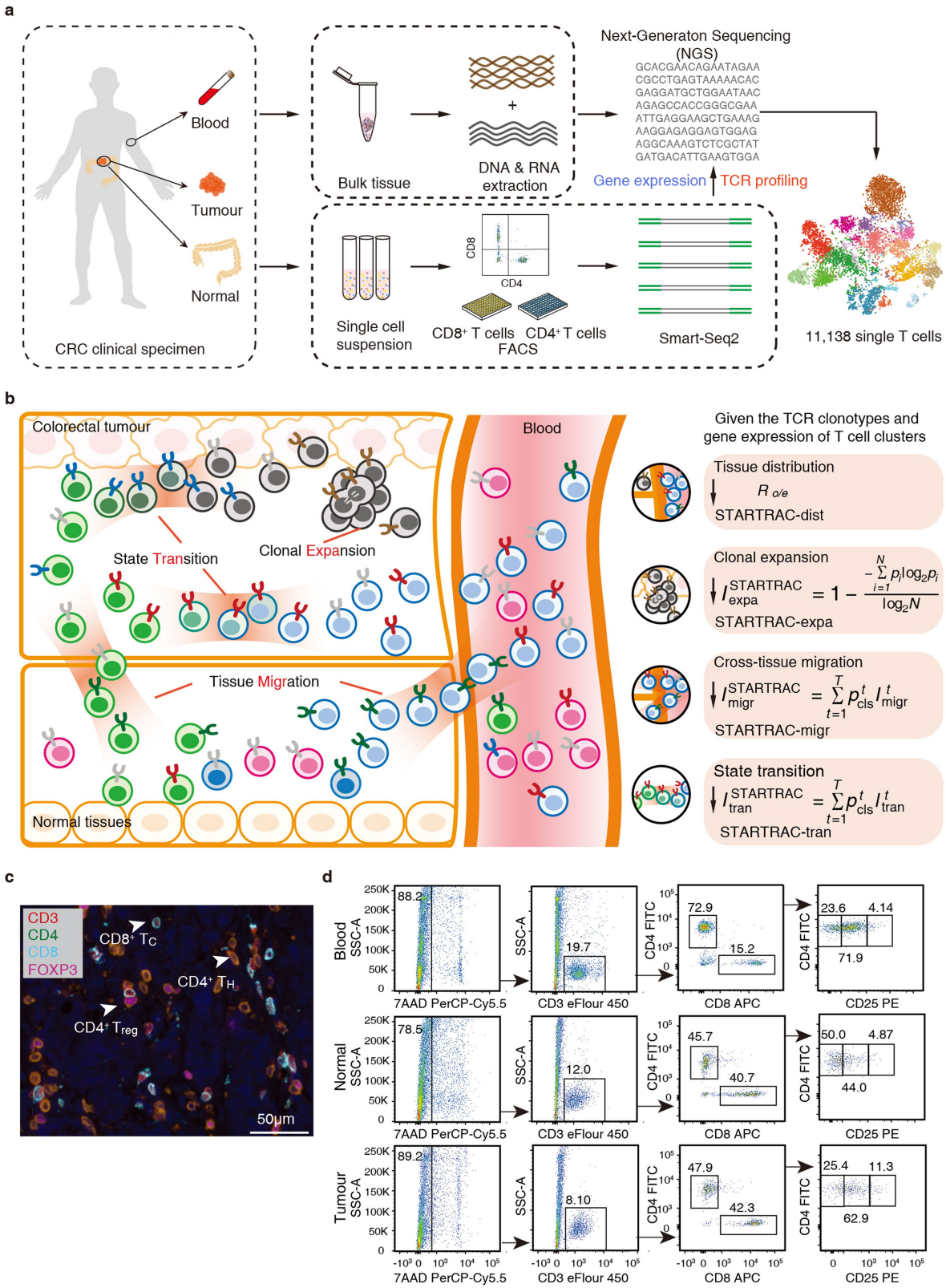
The data that support the findings of this study are available from the corresponding author upon request. Sequencing data are available at EGA (accession number EGAS00001002791), and processed gene expression data can be obtained from Gene Expression Omnibus (GEO) (accession number GSE108989).

31. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
32. Emtage, P. C. et al. Second-generation anti-carcinoembryonic antigen designer T cells resist activation-induced cell death, proliferate on tumor contact, secrete cytokines, and exhibit superior antitumor activity in vivo: a preclinical evaluation. *Clin. Cancer Res.* **14**, 8112–8122 (2008).

© 2018 Springer Nature Limited. All rights reserved.

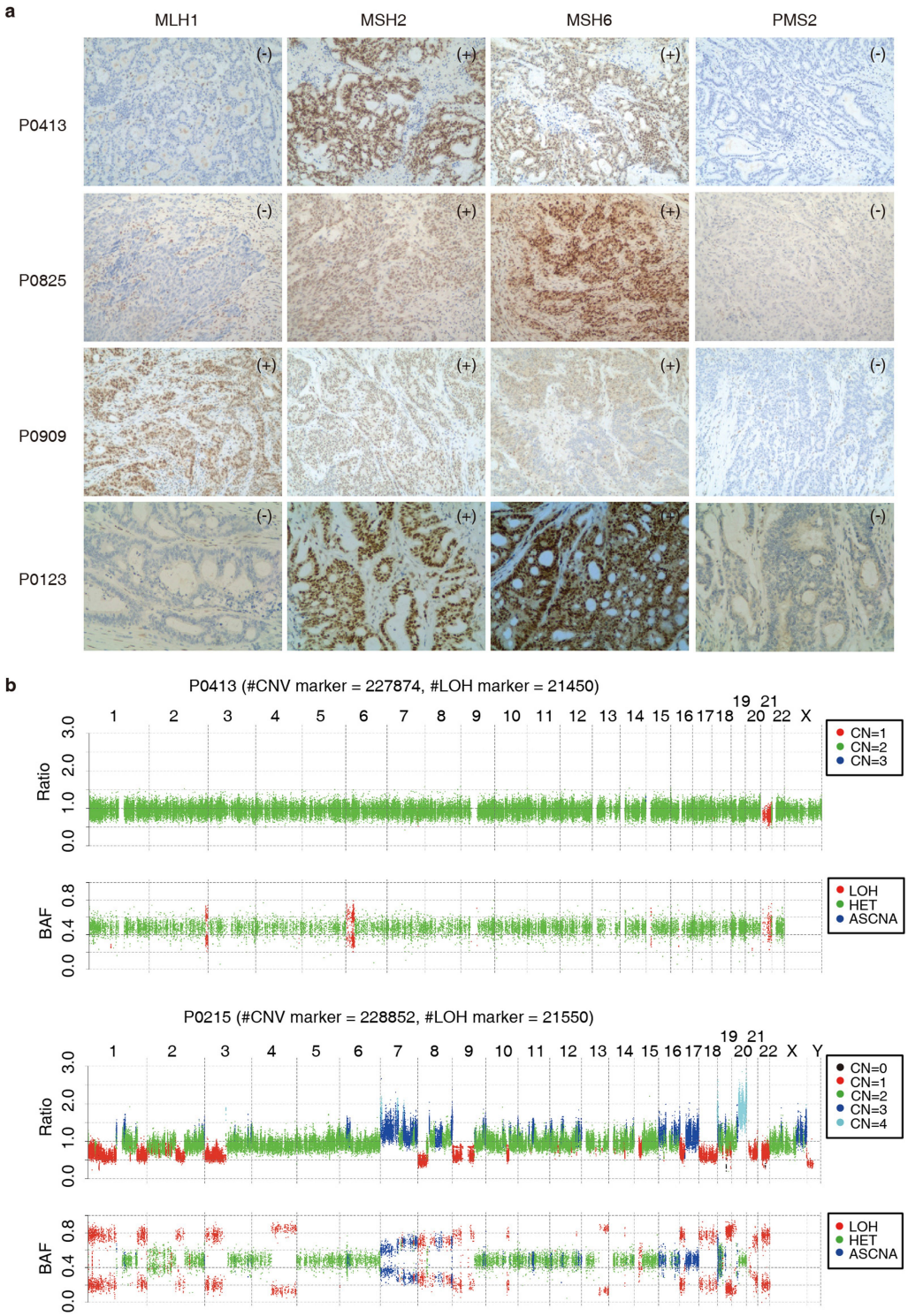
33. Chodisetti, S. B. et al. Triggering through Toll-like receptor 2 limits chronically stimulated T-helper type 1 cells from undergoing exhaustion. *J. Infect. Dis.* **211**, 486–496 (2015).
34. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
35. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
36. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
37. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
38. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
39. Amarasinghe, K. C. et al. Inferring copy number and genotype in tumour exome data. *BMC Genomics* **15**, 732 (2014).
40. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
41. van Wilgenburg, B. et al. MAIT cells are activated during human viral infections. *Nat. Commun.* **7**, 11653 (2016).
42. Godfrey, D. I., Stankovic, S. & Baxter, A. G. Raising the NKT cell family. *Nat. Immunol.* **11**, 197–206 (2010).
43. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
44. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
45. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
46. Strehl, A. & Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2002).
47. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
48. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
49. Whitfield, M. L., George, L. K., Grant, G. D. & Perou, C. M. Common markers of proliferation. *Nat. Rev. Cancer* **6**, 99–106 (2006).
50. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
51. Kirsch, I., Vignali, M. & Robins, H. T-cell receptor profiling in cancer. *Mol. Oncol.* **9**, 2063–2070 (2015).

© 2018 Springer Nature Limited. All rights reserved.



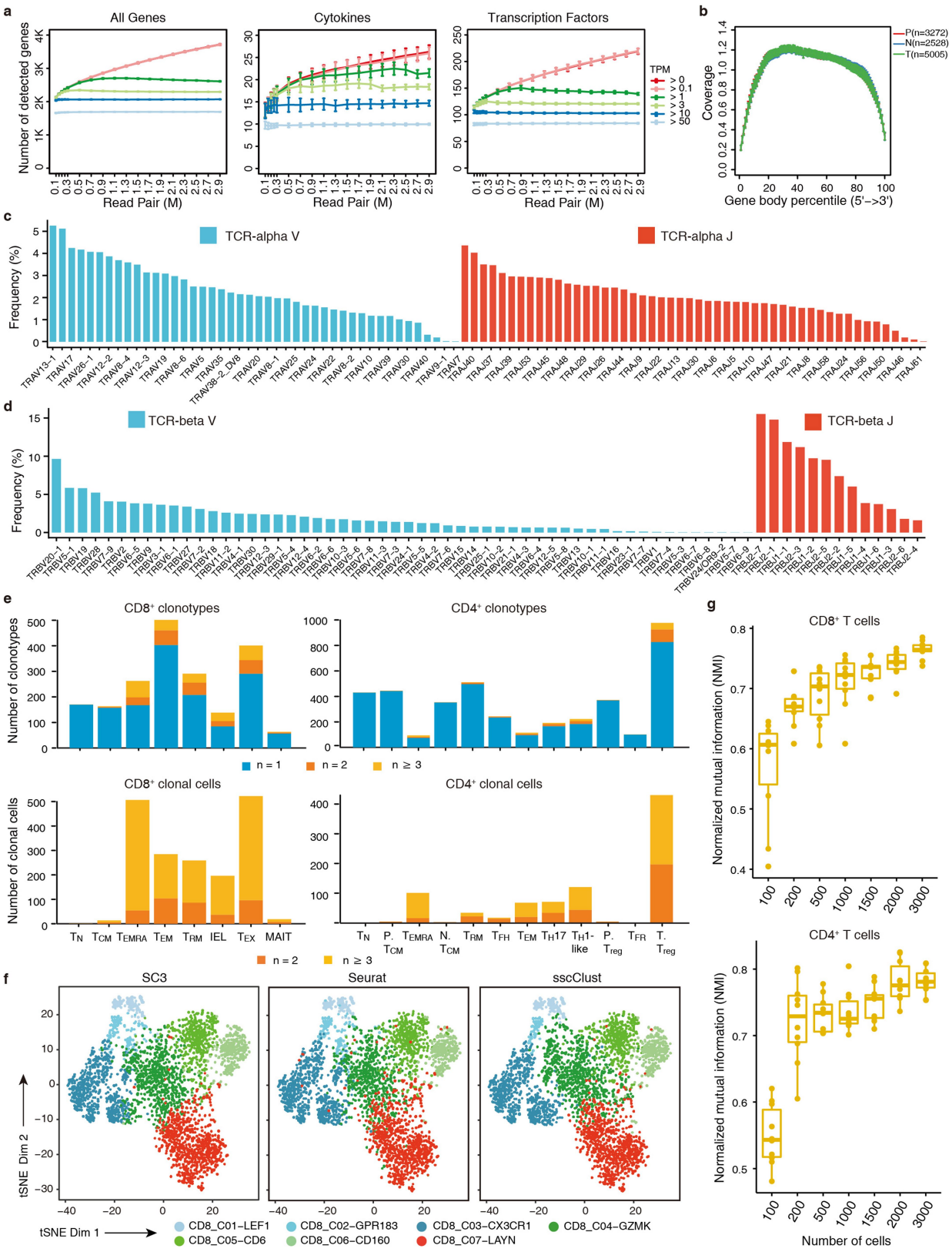
Extended Data Fig. 1 | Study design and tracking T cell dynamics of patients with CRC by STARTRAC. a, The experimental flowchart of this study. **b**, A cartoon illustrating four indices defined by STARTRAC to characterize T cell dynamics. STARTRAC-dist, tissue preference of a cluster estimated by ratios of observed cell numbers to random expectations ($R_{o/e}$); STARTRAC-expa, degree of clonal expansion of a cluster defined as '1 − evenness', with evenness as the normalized Shannon entropy of its TCR distribution; STARTRAC-migr, migratory potential of a cluster estimated by the average entropy of its clonotypes across tissues; STARTRAC-tran, potentials of developmental transitions of a cluster, estimated by the average entropy of its clonotypes across two different functional clusters. The detailed definitions of STARTRAC indices are in Methods. **c**, Opal multi-colour IHC staining with anti-CD3, -CD4, -CD8 and -FOXP3 antibodies to validate the existence of T cells in CRC tumours (exemplified by patient P0215). Original magnification, ×20. T_C, $CD8^+$ cytotoxic T cells; T_H, $CD4^+$ T helper cells. **d**, Gating strategy for single T cell sorting in this study (exemplified by patient P0215). T_C, T_H and T_{reg} cells were enriched by sorting $7AAD^-CD3^+CD8^+$, $7AAD^-CD3^+CD4^+$ $CD25^{-/int}$ and $7AAD^-CD3^+CD4^+$ $CD25^{hi}$ T cells, respectively.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | Pathological and genomic characteristics of CRC tumours in the study. a, Deficiency of mismatch repair proteins including MLH1, MSH2, MSH6 and PMS2 in all MSI patients (P0413, P0825, P0909 and P0123) measured by IHC ($n = 12$ patients). +, proficiency; −, deficiency. Original magnification, ×200. **b**, Profiles of DNA copy numbers of two representative patients (MSI patient, top; MSS patient, bottom). The copy number information was obtained by ADTex and depicted in bin count plots across chromosomes. The read count ratios ('1' in *y* axis means baseline copy number) and B allele frequencies (BAF) are shown. Various coloured dots in the ratio graph represent different copy number status of each segment. ASCNA, allele-specific copy number alteration; HET, heterozygous; LOH, loss of heterozygosity.

© 2018 Springer Nature Limited. All rights reserved.

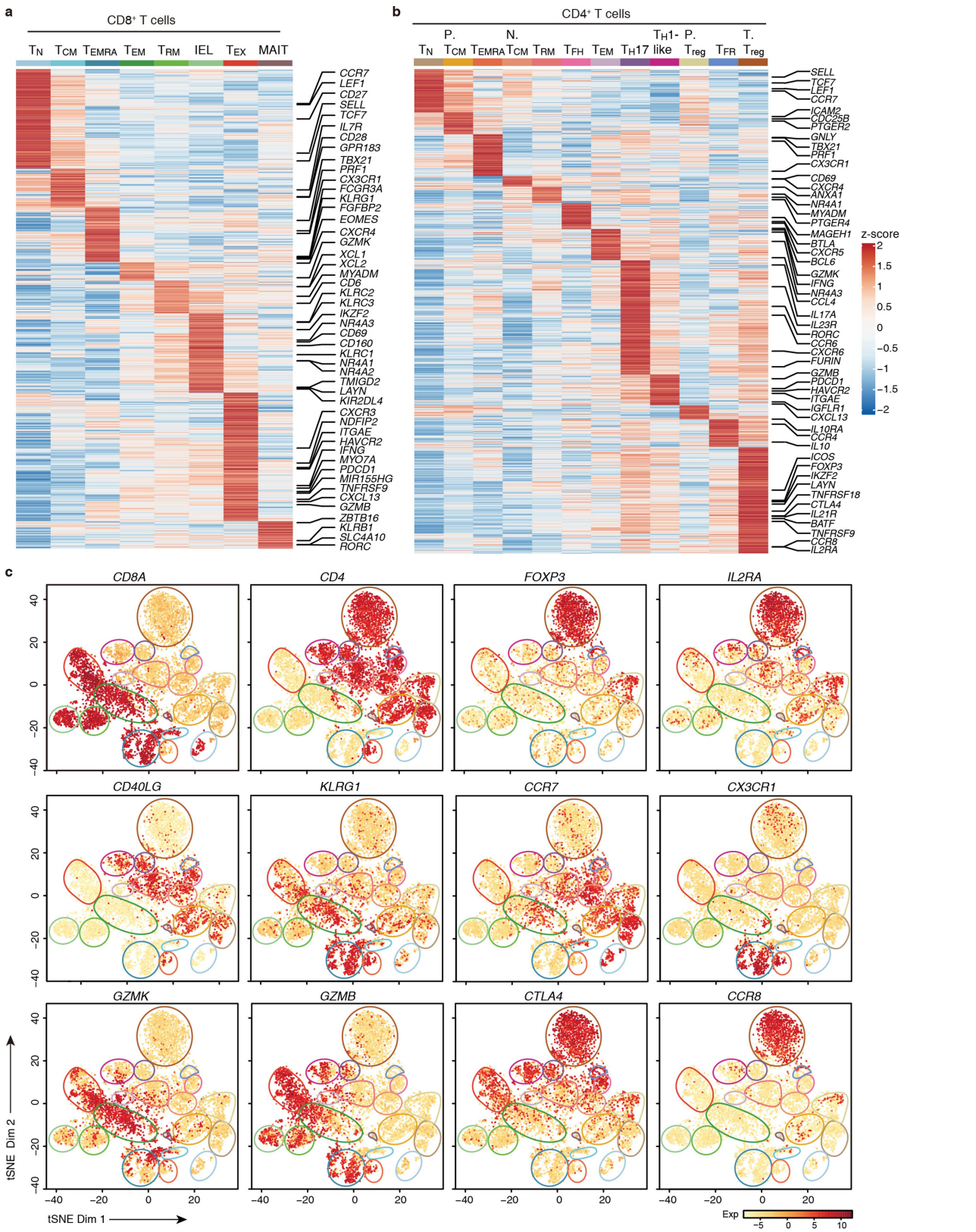


Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

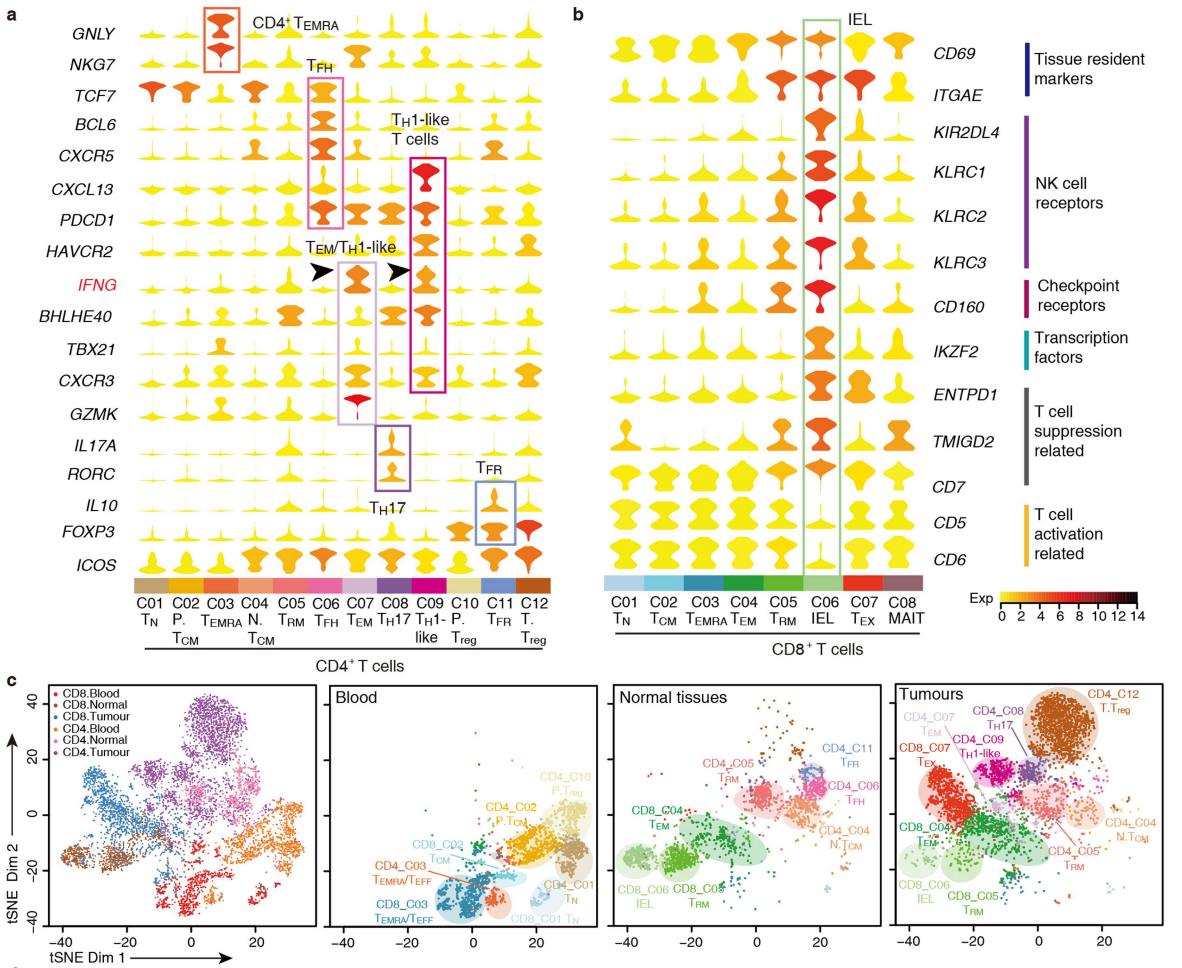
Extended Data Fig. 3 | Basic information of the single T cell RNA-seq data. a, Saturation curves of the number of detected genes against sequencing depth (exemplified by cell NTC-53 from patient P1228). Each point on the curve is derived from calculations based on the random selection of a fraction of raw reads from each sample, representing the average of 100 replicate sub-samplings. Error bars denote s.d. Each line with a different colour shows how fast a gene can reach detection saturation at different expression levels, represented by a particular TPM value. **b**, Unbiased coverage of gene body from 5′ to 3′ between blood, tumours and adjacent normal tissues. **c**, Frequencies of the V and J segments of the TCR α chains. **d**, Frequencies of the V and J segments of the TCR β chains. **e**, Bar plots showing the number of clonotypes and clonal cells in each $CD8^+$ and $CD4^+$ T cell cluster. The clonotypes are categorized as unique ($n = 1$) and clonal ($n = 2$ and $n \geq 3$) based on their cell numbers. Clonal cells are defined as those clonotypes containing at least two cells. **f**, *t*-SNE projection of 3,557 $CD8^+$ T cells (CD8_C01-LEF1, $n = 174$; CD8_C02-GPR183, $n = 169$; CD8_C03-CX3CR1, $n = 743$; CD8_C04-GZMK, $n = 773$; CD8_C05-CD6, $n = 487$; CD8_C06-CD160, $n = 351$; CD8_C07-LAYN, $n = 860$) based on different clustering methods including SC3, Seurat and sscClust. Each point represents one single cell coloured by cluster label. **g**. Box plots showing the down-sampling analysis of clustering performed on $CD8^+$ and $CD4^+$ T cell dataset. Each dot represents an individual clustering of a given number of T cells. The down-sampling and clustering were performed iteratively for each cell number ($n = 10$ times). Each down-sampled clustering was compared to the clustering performed on the entire dataset, using the NMI index. Higher NMI values indicate more accurate cluster assignment.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Expression levels of signature genes in each T cell cluster. a, Gene expression heat map of 8 CD8$^+$ T cell ($n = 3,628$) clusters. Rows represent signature genes and columns represent different clusters. **b**, Gene expression heat map of 12 CD4$^+$ T cell clusters ($n = 4,902$). **c**, *t*-SNE plot of expression levels of selected genes in different clusters indicated by the coloured oval corresponding to Fig. 1a. Number of cells contained in each cluster: CD8_C01-LEF1, $n = 174$; CD8_C02-GPR183, $n = 169$; CD8_C03-CX3CR1, $n = 743$; CD8_C04-GZMK, $n = 773$; CD8_C05-CD6, $n = 487$; CD8_C06-CD160, $n = 351$; CD8_C07-LAYN, $n = 860$; CD8_C08-SLC4A40, $n = 71$; CD4_C01-CCR7, $n = 462$; CD4_C02-ANXA1, $n = 472$; CD4_C03-GNLY, $n = 190$; CD4_C04-TCF7, $n = 388$; CD4_C05-CXCR6, $n = 568$; CD4_C06-CXCR5, $n = 262$; CD4_C07-GZMK, $n = 185$; CD4_C08-IL23R, $n = 244$; CD4_C09-CXCL13, $n = 319$; CD4_C10-FOXP3, $n = 389$; CD4_C11-IL10, $n = 103$; CD4_C12-CTLA4, $n = 1,320$.

© 2018 Springer Nature Limited. All rights reserved.



d

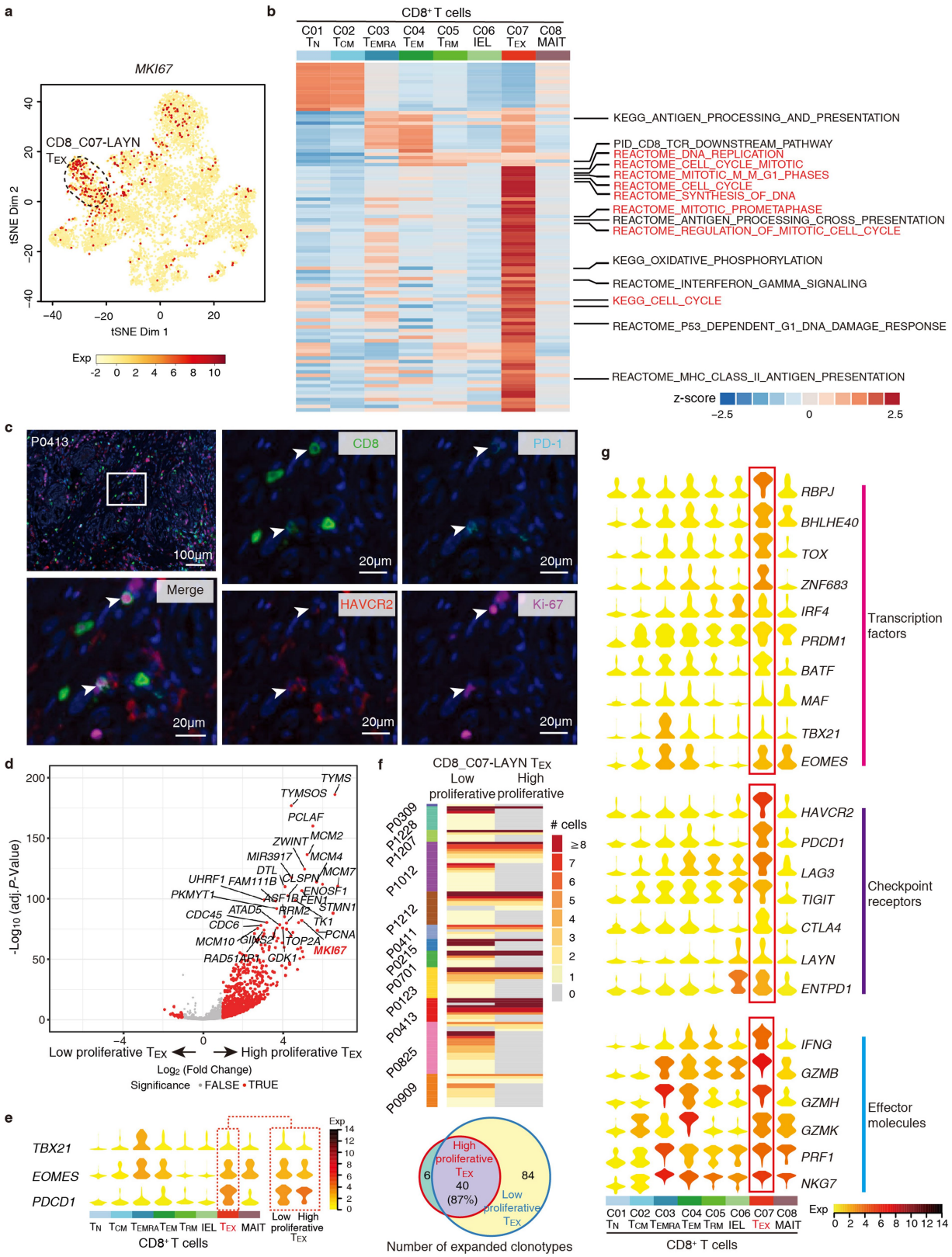
CRC T cell cluster names	Representative genes	Functional properties	STARTRAC-dist			STARTRAC-		
			P	N	T	expa	migr	tran
CD8_C01-LEF1	*CCR7, LEF1, SELL, TCF7, CD27, CD28, S1PR1*	CD8+ T_N	+++	+	+/-	-	-	-
CD8_C02-GPR183	*CCR7, SELL, IL7R, CD27, CD28, PRF1, GZMA, CCL5, GPR183, S1PR1*	CD8+ T_{CM}	+++	+/-	+	+	-	++
CD8_C03-CX3CR1	*KLRG1, CX3CR1, FCGR3A, FGFBP2, PRF1, GZMH, TBX21, EOMES, S1PR1,S1PR5*	CD8+ $T_{EMRA/EFF}$	+++	+/-	+/-	+++	+++	+++
CD8_C04-GZMK	*GZMK, CXCR4, CXCR3, CD44*	CD8+ T_{EM}	+	+++	+++	+	++	+++
CD8_C05-CD6	*CD6, XCL1, XCL2, MYADM, CAPG, RORA, NR4A1/2/3, CD69, ITGAE*	CD8+ T_{RM}	-	+++	+	++	+	+
CD8_C06-CD160	*CD160, KIR2DL4, TMIGD2, KLRC1/2/3, NR4A1/2/3, IKZF2, ENTPD1, CD69, ITGAE*	CD8+ IEL	+/-	+++	+	++	-	-
CD8_C07-LAYN	*HAVCR2, CXCL13, PDCD1, LAYN, TOX, IFNG, GZMB, MIR155HG, TNFRSF9, ITGAE*	CD8+ T_{EX}	-	+/-	+++	+++	-	+
CD8_C08-SLC4A10	*SLC4A10, KLRB1, ZBTB16, NCR3, RORC, RORA*	MAIT	+++	+	++	+	+	-
CD4_C01-CCR7	*CCR7, LEF1, SELL, TCF7, CD27, CD28, S1PR1*	CD4+ T_N	+++	+/-	+/-	-	-	-
CD4_C02-ANXA1	*CCR7, SELL, PTGER2, ICAM2, ANXA1, ANXA2, S1PR1*	CD4+ Blood-T_{CM}	+++	+/-	+/-	-	-	-
CD4_C03-GNLY	*KLRG1, CX3CR1, NKG7, PRF1, GNLY, GZMH, TBX21, CTSW, S1PR1, S1PR5*	CD4+ $T_{EMRA/EFF}$	+++	+/-	+/-	+++	++	+++
CD4_C04-TCF7	*CCR7, TCF7, RGS1, CD69*	CD4+ Normal-T_{CM}	+/-	+++	+	-	-	-
CD4_C05-CXCR6	*CD69, KLRB1, PTGER4, IL7R, CXCR6, NR4A1/2/3, MYADM*	CD4+ T_{RM}	+/-	+++	+++	-	-	-
CD4_C06-CXCR5	*CXCR5, BCL6, ICA1, TOX, TOX2, IL6ST, MAGEH1, BTLA, ICOS, PDCD1, CD200*	T_{FH}	+/-	+++	+	-	-	-
CD4_C07-GZMK	*GZMK, GZMA, CCL5, IFNG, RUNX3, EOMES, CXCR3, CXCR4, CD44*	CD4+ T_{EM}/T_H1-like	+/-	+	+++	++	+	++
CD4_C08-IL23R	*IL23R, RORC, IL17A, FURIN, CTSH, CCR6, KLRB1, CAPG, ITGAE*	T_H17	-	+	+++	+	-	+
CD4_C09-CXCL13	*CXCL13, IFNG, CXCR3, BHLHE40, GZMB, PDCD1, HAVCR2, ICOS, IGFLR1, ITGAE*	T_H1-like T cells	-	+/-	+++	+	-	+
CD4_C10-FOXP3	*FOXP3, IL2RA, IL10RA, IKZF2, RTKN2, CDC25B, S1PR4*	Blood-T_{reg}	+++	+	+/-	-	-	-
CD4_C11-IL10	*FOXP3, IL2RA, CXCR5, PDCD1, IL10, CCR4, CD69*	T_{FR}	-	+++	+	-	-	-
CD4_C12-CTLA4	*FOXP3, CCR8, TNFRSF18, LAYN, TNFRSF9, IKZF2, RTKN2, CTLA4, BATF, IL21R*	Tumour-T_{reg}	-	+	+++	+	-	-

Extended Data Fig. 5 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 5 | Summary of functional properties of various T cell clusters. a, Functional subsets of $CD4^+$ T cell ($n = 4{,}902$) clusters defined by a set of known marker genes. Number of cells contained in each $CD4^+$ cluster: T_N, $n = 462$; $P.T_{CM}$, $n = 472$; T_{EMRA}, $n = 190$; $N.T_{CM}$, $n = 388$; T_{RM}, $n = 568$; follicular T helper (T_{FH}), $n = 262$; T_{EM}, $n = 185$; T_H17, $n = 244$; T_H1-like cells, $n = 319$; $P.T_{reg}$, $n = 389$; $N.T_{reg}$, $n = 103$; $T.T_{reg}$, $n = 1{,}320$. N, normal tissue; P, peripheral blood; T, tumour. **b**, Characteristics of the $CD8^+$ IEL T cells as defined by the expression properties of a panel of functionally relevant genes in $CD8^+$ T cells ($n = 3{,}628$). Number of cells contained in each $CD8^+$ cluster: T_N, $n = 174$; T_{CM}, $n = 169$; T_{EMRA}, $n = 743$; T_{EM}, $n = 773$; T_{RM}, $n = 487$; IEL, $n = 351$; T_{EX}, $n = 860$; MAIT, $n = 71$. For violin plots in **a** and **b**, colours denote average expression levels; widths denote cell densities. **c**, *t*-SNE plot showing the presence of different T cell clusters in peripheral blood ($n = 2{,}449$; $CD8^+$ T cells, $n = 1{,}021$; $CD4^+$ T cells, $n = 1{,}428$), adjacent normal tissues ($n = 1{,}962$; $CD8^+$ T cells, $n = 961$; $CD4^+$ T cells, $n = 1{,}001$) and tumours ($n = 4{,}119$; $CD8^+$ T cells, $n = 1{,}646$; $CD4^+$ T cells, $n = 2{,}473$). **d**, Overview of T cell cluster characteristics. STARTRAC-dist: +++ indicates $R_{o/e} > 1$; ++, $0.8 < R_{o/e} \leq 1$; +, $0.2 < R_{o/e} \leq 0.8$; +/−, $0 < R_{o/e} \leq 0.2$; −, $R_{o/e} = 0$. STARTRAC-expa: +++ indicates $I_{expa}^{STARTRAC} > 0.10$; ++, $0.06 < I_{expa}^{STARTRAC} \leq 0.10$; +, $0.005 < I_{expa}^{STARTRAC} \leq 0.05$; -, $I_{expa}^{STARTRAC} \leq 0.005$. STARTRAC-migr: +++ indicates $I_{migr}^{STARTRAC} > 0.50$; ++, $0.21 < I_{migr}^{STARTRAC} \leq 0.50$; +, $0.1 < I_{migr}^{STARTRAC} \leq 0.20$; −, $I_{migr}^{STARTRAC} \leq 0.1$. STARTRAC-tran, +++ indicates $I_{tran}^{STARTRAC} > 0.20$; ++, $0.10 < I_{tran}^{STARTRAC} \leq 0.20$; +, $0.05 < I_{tran}^{STARTRAC} \leq 0.10$; −, $I_{tran}^{STARTRAC} \leq 0.05$.

© 2018 Springer Nature Limited. All rights reserved.

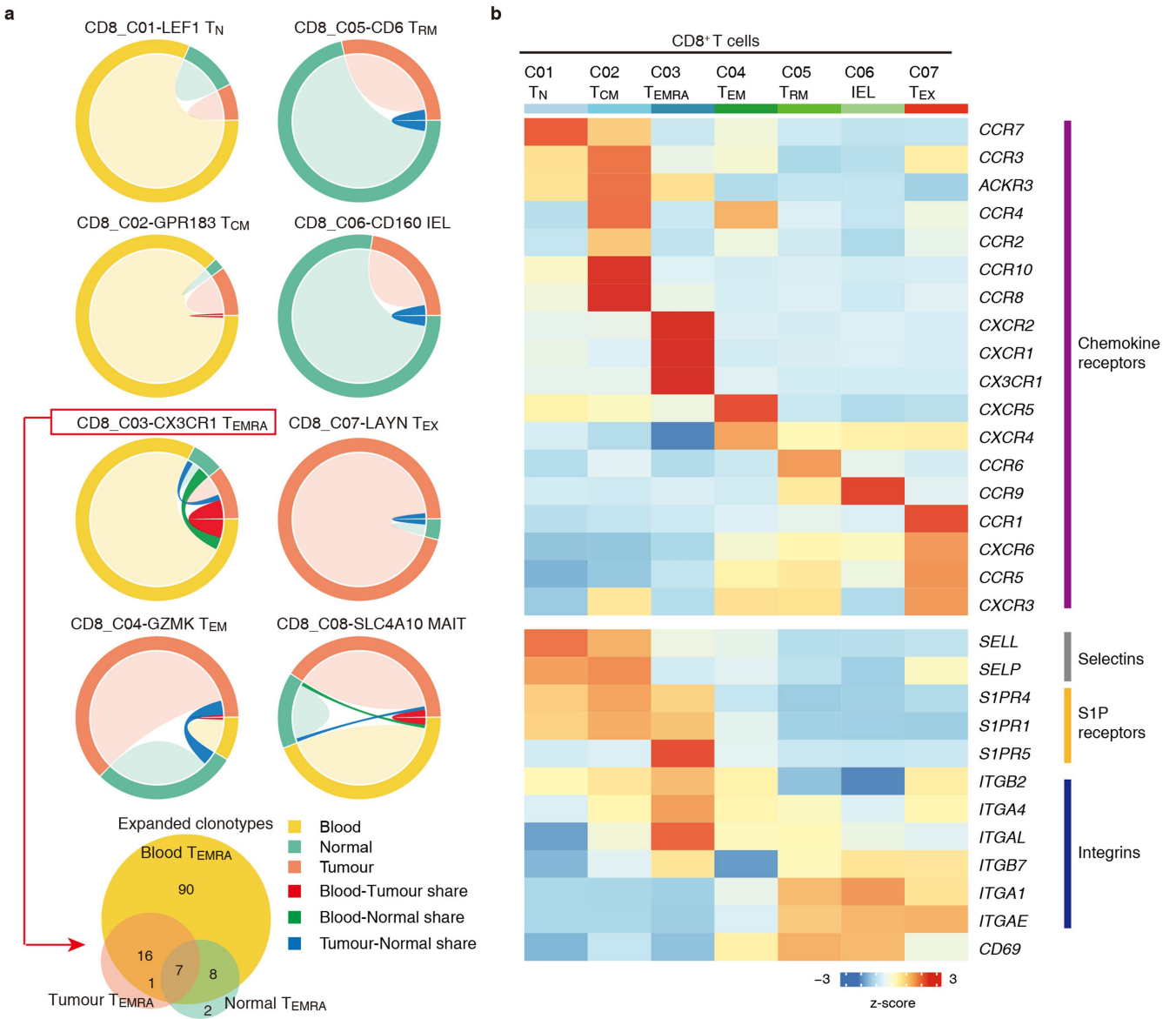


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

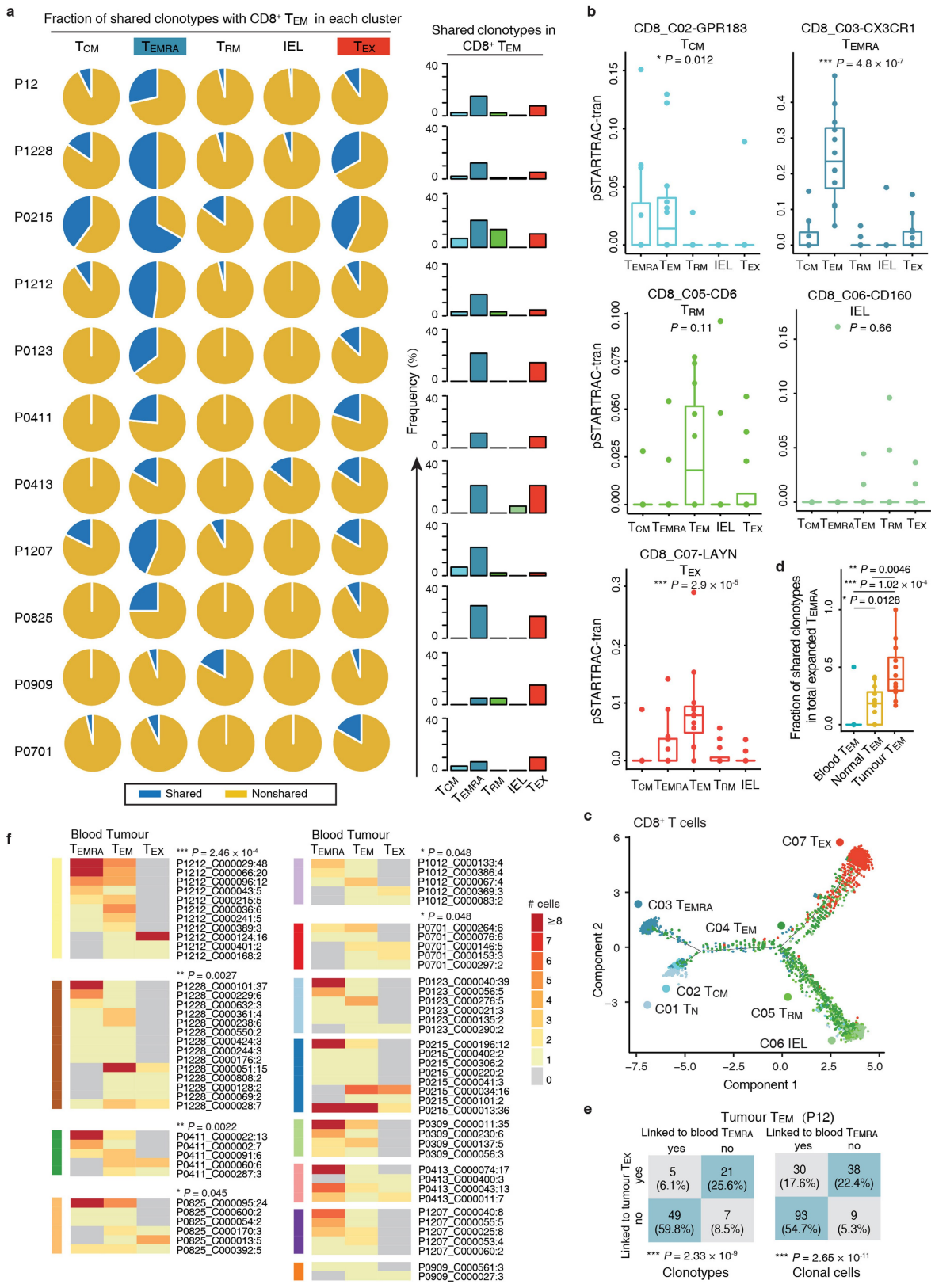
Extended Data Fig. 6 | $CD8^+$ T_{EX} cells are characterized by high proliferation property and production of effector molecules. a, A subpopulation of $CD8^+$ T_{EX} shows high expression of *MKI67* among 8,530 T cells. **b**, Gene set enrichment analysis (GSEA) showing the enrichment of proliferation-related pathways in $CD8^+$ T_{EX} cells ($n = 3{,}628$; false discovery rate < 0.01; labelled in red). **c**, Representative example of a CRC tumour stained by multi-coloured IHC showing co-expression of Ki67, CD8, PD-1 and HAVCR2 in $CD8^+$ T_{EX} cells (exemplified by P0413; $n = 2$ patients). Original magnification, $\times 20$. **d**, Volcano plot showing the differentially expressed genes between high-proliferative ($n = 140$) and low-proliferative ($n = 720$) T_{EX} cells. Most of the highly expressed genes in high-proliferative T_{EX} cells are related to cell proliferation. Adjusted $P < 0.01$; fold change ≥ 2; two-sided unpaired limma-moderated *t*-test; Benjamini–Hochberg adjusted *P* value **e**, Violin plot showing the expression of *TBX21*, *EOMES* and *PDCD1* in each $CD8^+$ T cell ($n = 3{,}628$) cluster and the low-proliferative ($n = 720$) or high-proliferative ($n = 140$) T_{EX} cell subsets. **f**, Most of the clonotypes of high-proliferative T_{EX} cells were also found in low-proliferative T_{EX} cells (top). Each row represents an individual clonotype from one patient. Venn diagram showing overlapped clonal clonotypes (≥ 2 cells) of high- and low-proliferative T_{EX} cells (bottom). **g**. Characteristics of $CD8^+$ T_{EX} cells ($n = 3{,}628$) as defined by the gene expression of a series of transcription factors, checkpoint receptors, and effector molecules. For violin plots in **e** and **g**, colours denote average expression levels; widths denote cell densities.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Distinct migration capabilities of different CD8$^+$ T cell clusters. a, Top, chord diagram showing the distribution of clonotypes in blood, normal mucosa and tumours for different CD8$^+$ T cell clusters. CD8$^+$ T_{EMRA} cells show remarkable TCR sharing among different tissues. The shadows coloured in transparent yellow, green and orange represent blood, normal and tumour-specific clonotypes, respectively. The bridges coloured in dark green, dark red and dark blue represent clonotypes shared by blood–normal, blood–tumour and normal–tumour, respectively. Bottom, Venn diagram showing the distribution of expanded clonotypes in blood, tumour and normal T_{EMRA} cells. **b**, Relative average expression patterns of migration-related genes across CD8$^+$ T cell clusters (total $n = 3{,}557$ cells, excluding MAIT cells). Number of cells contained in each cluster: T_N, $n = 174$; T_{CM}, $n = 169$; T_{EMRA}, $n = 743$; T_{EM}, $n = 773$; T_{RM}, $n = 487$; IEL, $n = 351$; T_{EX}, $n = 860$.

© 2018 Springer Nature Limited. All rights reserved.

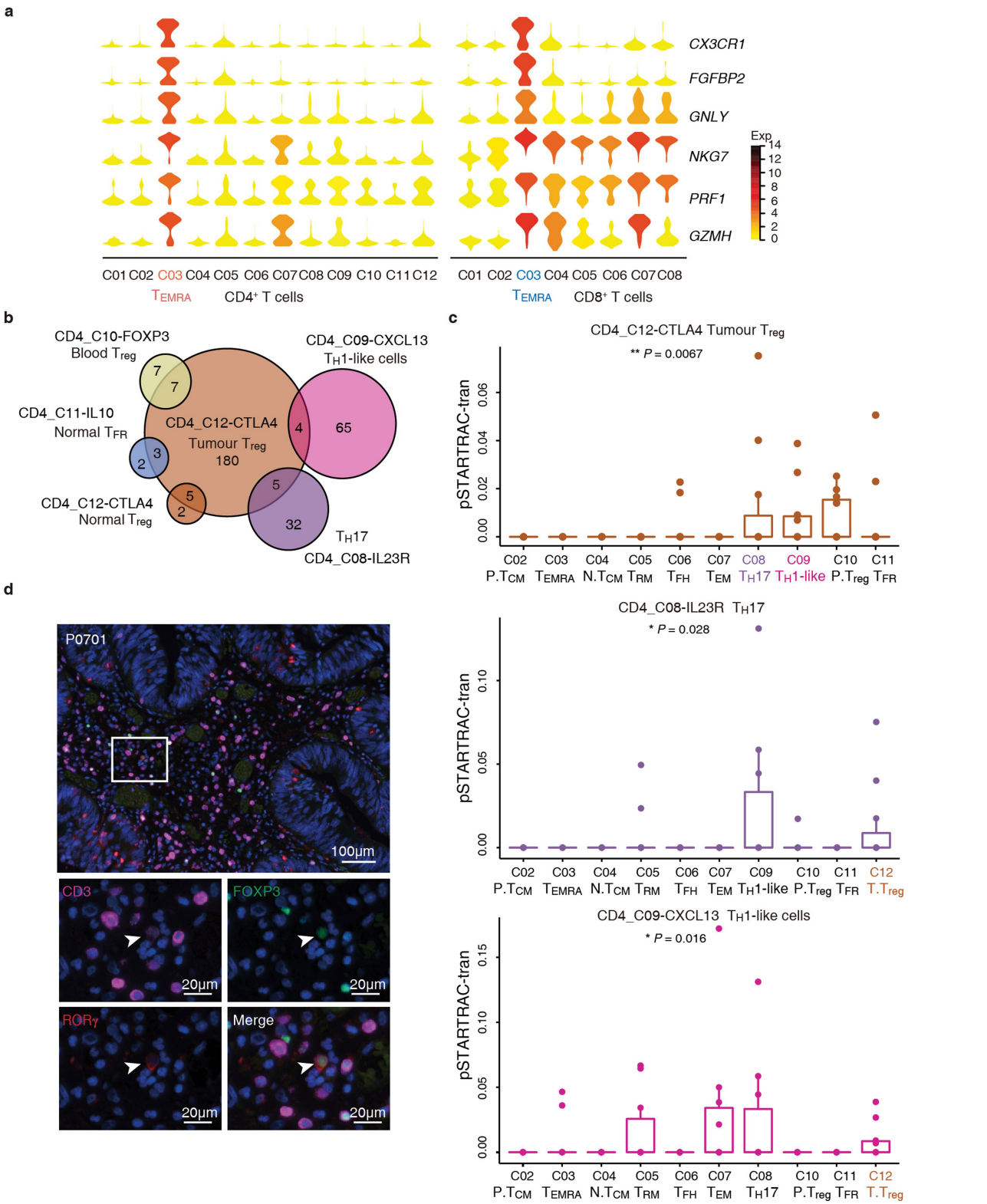


Extended Data Fig. 8 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

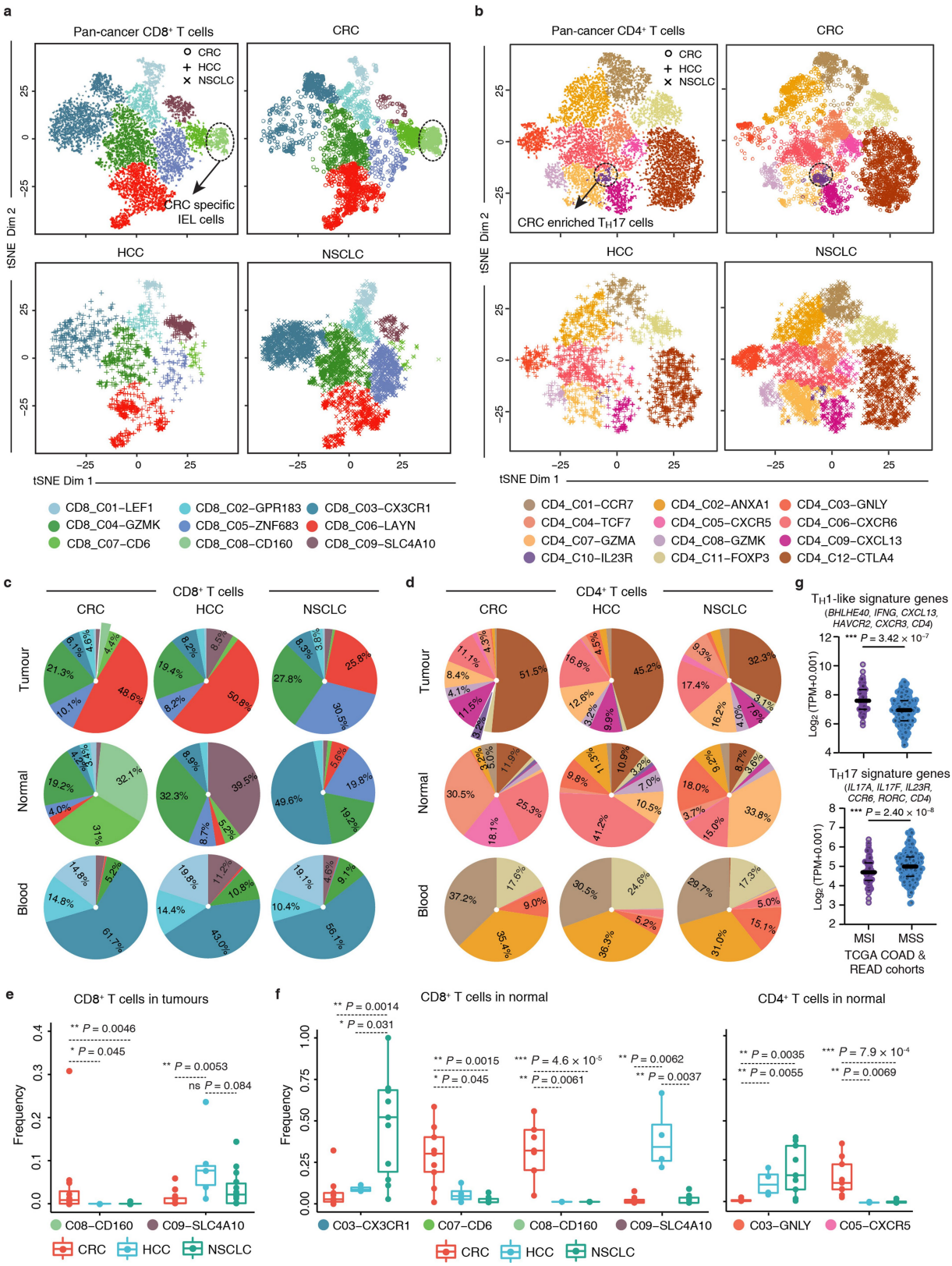
Extended Data Fig. 8 | TCR sharing and state transitions of CD8$^+$ T cell clusters implicated by STARTRAC-tran indices. a, Pie charts showing the fraction of shared clonotypes with CD8$^+$ T_{EM} cells within the other indicated clusters (left). P12 represents merged data of 12 patients with CRC. Bar plots showing the fraction of shared clonotypes of CD8$^+$ T_{EM} with other clusters within the CD8$^+$ T_{EM}. **b**, pSTARTRAC-tran indices of CD8$^+$ T_{CM}, T_{EMRA}, T_{RM}, IEL and T_{EX} cells for each patient (depicted by dots). *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Kruskal–Wallis test. **c**, Potential developmental trajectory of CD8$^+$ T cells ($n = 3{,}557$, excluding MAIT cells) inferred by Monocle2 based on gene expressions. **d**, Frequency of shared clonotypes in CD8$^+$ T_{EMRA} cells with various T_{EM} cell subsets in each patient ($n = 12$). **e**, Statistical analysis of tumour T_{EM} shared TCRs with blood T_{EMRA} and tumour T_{EX} cells based on the number of clonotypes and clonal cells (related to Fig. 1h). ***$P < 0.001$, two-sided Fisher's exact test. **f**, Clonotypes of tumour T_{EM} cells crossing different clusters showing mutually exclusive TCR sharing of tumour T_{EM} cells with blood T_{EMRA} and tumour T_{EX} cells. Each row represents an individual clonotype from one patient. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Fisher's exact test (based on the number of clonal cells in each patient). Number of clonal cells analysed in each patient: P1212, $n = 30$; P1228, $n = 27$; P0411, $n = 11$; P0825, $n = 10$; P1012, $n = 7$; P0701, $n = 9$; P0123, $n = 9$; P0215, $n = 17$; P0309, $n = 9$; P0413, $n = 2$; P1207, $n = 7$; P0909, $n = 2$.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Characterization of CD4$^+$ T_{EMRA} and tumour T_{reg} cells by STARTRAC analysis. a, Violin plots showing normalized expression of cytotoxic related molecules in 12 CD4$^+$ ($n = 4{,}902$ cells) and 8 CD8$^+$ ($n = 3{,}628$) T cell clusters. Colours denote mean values; width denotes cell densities. **b**, Venn diagram highlighting common clonotypes ($n_{cell} \geq 2$) shared between tumour T_{reg} and other CD4$^+$ T cell clusters. **c**, Developmental transition of tumour T_{reg} cells, T_H17 cells and T_H1-like cells with other CD4$^+$ cells quantified by pSTARTRAC-tran indices for each patient ($n = 11$). **d**, Representative example of a CRC tumour stained by IHC, with white arrow showing co-expression of CD3, FOXP3 and RORγ ($n = 2$ patients). Original magnification, ×20.

© 2018 Springer Nature Limited. All rights reserved.

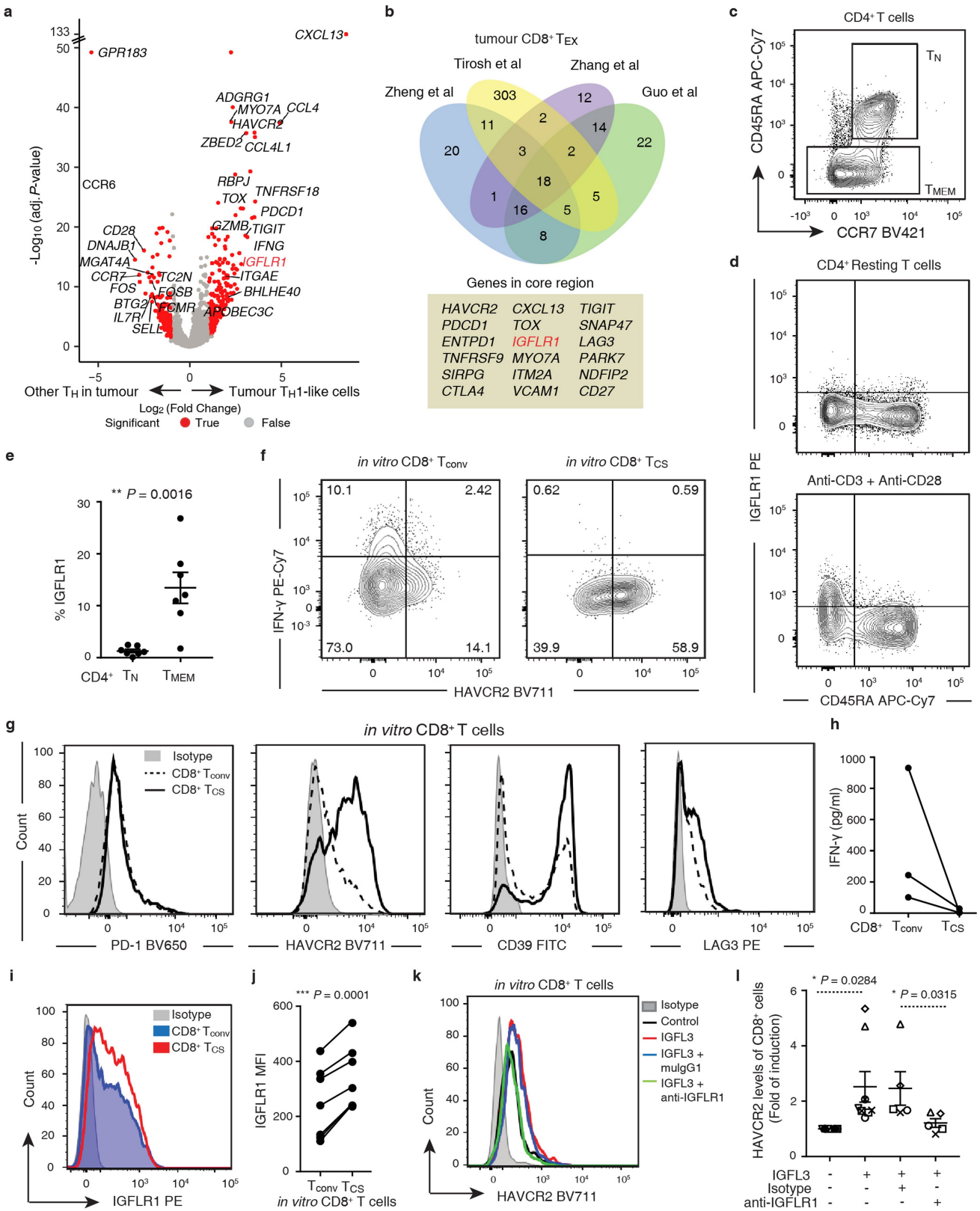


Extended Data Fig. 10 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | Comparative analysis of T cells from different cancer indications based on integrated analyses. a, *t*-SNE plot of 8,874 single $CD8^+$ T cells from this CRC study ($n = 3,632$), and previous HCC[9] ($n = 1,467$) and NSCLC[10] ($n = 3,775$) studies. Nine $CD8^+$ clusters were generated by sscClust based on the integrated dataset. The CRC-specific IEL cells (CD8_C06-CD160) are highlighted. **b**, *t*-SNE plot of 12,635 single $CD4^+$ T cells from this CRC study ($n = 4,929$), and previous HCC[9] ($n = 2,472$) and NSCLC[10] ($n = 5,234$) studies. The CRC-enriched T_H17 cells (CD4_C10-IL23R) are highlighted. Each dot represents one single cell coloured by clusters and shaped by tumour types in **a** and **b**. **c**, Composition of different $CD8^+$ T cells in each tumour type by different tissue origins. $CD8^+$ T cell clusters with frequencies below 3% are not labelled. **d**, Composition of different $CD4^+$ T cells in each tumour type by different tissue origins. $CD4^+$ T cell clusters with frequencies below 3% are not labelled. **e**, Comparison of the fractions of $CD8^+$ IEL (CD8_C08-CD160) and MAIT (CD8_C09-SLC4A10) cells in tumours from patients with CRC ($n = 12$), HCC ($n = 5$) and NSCLC ($n = 14$). **f**, Comparison of the fractions of different $CD8^+$ T cells and $CD4^+$ T cells in control tissues from patients with CRC ($n = 12$), HCC ($n = 5$) and NSCLC ($n = 14$). **g**, Validation of the enrichments of $CXCL13^+BHLHE40^+$ T_H1-like cells in patients with MSI-H CRC ($n = 62$) and T_H17 cells in patients with MSS CRC ($n = 286$) in the TCGA COAD and READ cohorts by comparison of the indicated signature gene expression. Centre lines denote the median, top and bottom lines denote the 25th and 75th percentiles. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; two-sided Wilcoxon test (**e**–**g**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 11 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 11 | IGFLR1 expression in activated CD4$^+$ T cells and exhausted CD8$^+$ T cells. a, Volcano plot showing differentially expressed genes between tumour *CXCL13*$^+$*BHLHE40*$^+$ T_H1-like T cells ($n = 203$) and other T_H cells in tumours ($n = 723$; Supplementary Table 10). Adjusted $P < 0.01$ (two-sided unpaired limma-moderated *t*-test; Benjamini–Hochberg adjusted *P* value) and fold change ≥ 2. **b**, Venn diagram showing the overlap of tumour CD8$^+$ exhaustion-related genes identified in this study ($n = 68$, Supplementary Table 11) with those from previous melanoma[8] ($n = 349$), HCC[9] ($n = 82$) and NSCLC[10] ($n = 90$) studies. The detailed overlaps of CD8$^+$ exhaustion-related genes in different cancer types are in Supplementary Table 11. $P < 2.2 \times 10^{-16}$, hypergeometric test. **c**, CD4$^+$ naive (T_N) and memory (T_{MEM}) T cells were gated as CD45RA$^+$CCR7$^+$ and CD45RA$^-$CCR7$^{+/-}$ cells by FACS. **d**, FACS plots of IGFLR1 expression in activated CD4$^+$ T cells ($n = 6$ donors, $n = 3$ independent experiments). **e**, Quantification of IGFLR1 expression levels from **d** as a percentage of IGFLR1$^+$ T_N or T_{MEM} CD4$^+$ subsets under suboptimal activation conditions ($n = 7$). Each symbol represents a donor with mean ± s.e.m. shown (**e**, **l**). **f**, Representative FACS plots for HAVCR2 and IFNγ expression levels in CD8$^+$ T_{conv} (activated by anti-CD3 plus anti-CD28) and T_{CS} cells (in vitro chronically stimulated exhausted CD8$^+$ T cells from corresponding individuals). Numbers in quadrants indicate the percentage of positive cells ($n = 5$ donors, $n = 2$ independent experiments). **g**, Representative histograms of PD-1, HAVCR2 ($n = 8$ donors, $n = 3$ independent experiments), CD39 and LAG3 ($n = 4$ donors, $n = 2$ independent experiments) expression levels in CD8$^+$ T_{conv} and T_{CS} cells. **h**, Quantification of IFNγ levels produced by CD8$^+$ T_{conv} and T_{EX} cells from **g** of three donors. **i**, Representative histogram of IGFLR1 expression levels in CD8$^+$ T_{conv} and T_{CS} cells. **j**, Expression levels of IGFLR1 in activated CD8$^+$ T_{conv} and T_{CS} cells determined by FACS (MFI, mean fluorescent intensity; $n = 6$ donors, $n = 4$ independent experiments). **k**, Representative histograms of HAVCR2 expression in T_{CS} cells subjected to re-stimulation with anti-CD3 alone (control) or together with recombinant human IGFL3 as well as indicated antibodies for 2 days ($n = 5$ donors, $n = 3$ independent experiments). **l**, Quantification of HAVCR2 levels from **k**. Two-sided paired Student's *t*-test (**e**, **j** and **l**).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0774-y

Disruption of a self-amplifying catecholamine loop reduces cytokine release syndrome

Verena Staedtke[1,2,7]*, Ren-Yuan Bai[3,7]*, Kibem Kim[1], Martin Darvas[4], Marco L. Davila[5], Gregory J. Riggins[3], Paul B. Rothman[6], Nickolas Papadopoulos[1], Kenneth W. Kinzler[1], Bert Vogelstein[1]* & Shibin Zhou[1]*

Cytokine release syndrome (CRS) is a life-threatening complication of several new immunotherapies used to treat cancers and autoimmune diseases[1–5]. Here we report that atrial natriuretic peptide can protect mice from CRS induced by such agents by reducing the levels of circulating catecholamines. Catecholamines were found to orchestrate an immunodysregulation resulting from oncolytic bacteria and lipopolysaccharide through a self-amplifying loop in macrophages. Myeloid-specific deletion of tyrosine hydroxylase inhibited this circuit. Cytokine release induced by T-cell-activating therapeutic agents was also accompanied by a catecholamine surge and inhibition of catecholamine synthesis reduced cytokine release in vitro and in mice. Pharmacologic catecholamine blockade with metyrosine protected mice from lethal complications of CRS resulting from infections and various biotherapeutic agents including oncolytic bacteria, T-cell-targeting antibodies and CAR-T cells. Our study identifies catecholamines as an essential component of the cytokine release that can be modulated by specific blockers without impairing the therapeutic response.

Inflammation is crucial for immune defence against pathogens. However, when dysregulated, the cytokines that normally mediate protective immunity and promote recovery can cause a harmful systemic hyperactivated immune state known as cytokine release syndrome (CRS), which can lead to cardiovascular collapse, multiple organ dysfunction and death[1]. In addition to following infections by naturally occurring pathogens, CRS can be observed after biotherapeutic agents are administered to patients or to experimental animals, thereby seriously limiting the utility of these otherwise promising agents, which include oncolytic viruses and bacteria[3,6], recombinant lymphokines[5], natural and bispecific antibodies[2], and T cells designed to kill cancer cells[4].

The present study began with experiments employing the anaerobic spore-forming bacterial strain *Clostridium novyi*-NT to treat cancer[6]. *C. novyi*-NT spores germinate exclusively in hypoxic tumour tissues and can destroy them[6]. However, when high doses of spores were injected into very large tumours, a massive infection occurred and animals died within a few days with severe cytokine release due to a combination of tumour lysis and direct toxic effects of the bacteria (sepsis)[6] that was not reversed by the antibiotic metronidazole (Extended Data Fig. 1a). To mitigate this dose-limiting toxicity, we attempted to pre-treat mice with agents known to downregulate the inflammatory response[7,8]. Unfortunately, the anti-inflammatory agent dexamethasone and antibodies against TNF, IL-6 receptor (IL-6R) or IL-3 had limited effects on survival with only IL-6R blockade resulting in a significant but marginal improvement (Extended Data Fig. 1a).

We then engineered *C. novyi*-NT to secrete a number of anti-inflammatory proteins that might mitigate the bacteria-associated toxicity; atrial natriuretic peptide (ANP) was the only one that proved successful without compromising tumour lysis. ANP is an endogenous peptide released by cardiac cells, and regulates fluid and electrolyte homeostasis[9]. It has also been shown to have anti-inflammatory properties through reduction of cytokine release induced by lipopolysaccharide (LPS) in mice[9]. To investigate whether ANP could protect mice from severe bacterial infections such as those caused by *C. novyi*-NT, we engineered *C. novyi*-NT to express and secrete ANP by stably integrating an expression cassette of ANP with a signal peptide into the *C. novyi*-NT genome using the group II intron targeting[10]. Selected *C. novyi*-NT clones were characterized for ANP expression, biological activity and growth patterns in vitro (Extended Data Fig. 1b–d), and the clone with the highest expression of ANP (called 'ANP-*C. novyi*-NT', 1-29) was studied further.

One dose of ANP-*C. novyi*-NT spores injected into subcutaneously implanted CT26 tumours resulted in robust germination and tumour regression. Plasma levels of ANP and cyclic GMP (cGMP) in mice injected with ANP-*C. novyi*-NT were increased two to four times over that of mice injected with *C. novyi*-NT (Extended Data Fig. 1e, f). Strikingly, at similar efficiencies of germination and proliferation between the two strains (Extended Data Fig. 1g), more than 80% of the animals that received ANP-*C. novyi*-NT survived, 84% of which had complete tumour regression, whereas none of the mice treated with *C. novyi*-NT survived (Fig. 1a, upper and lower panel).

Mice injected with ANP-*C. novyi*-NT exhibited a noticeable reduction in tissue damage and inflammation. There were fewer infiltrating $CD11b^+$ myeloid cells in the liver, spleen and lungs (Fig. 1b, Extended Data Fig. 1h) and significant reductions in myeloid-derived cytokines (IL-1β, IL-6, MIP-2), chemoattractants (KC), and to a lesser degree, TNF and IFN-γ compared to mice treated with *C. novyi*-NT (Fig. 1c, Extended Data Fig. 1j). The latter cohort was also found to have an increased pulmonary permeability index (Extended Data Fig. 1i) and bone marrow myeloid hyperplasia (Fig. 1b, Extended Data Fig. 1h). Of interest, the diminished inflammatory response in mice treated with ANP-*C. novyi*-NT-treated was accompanied by markedly lower levels of circulating catecholamines (adrenaline, noradrenaline and dopamine) (Fig. 1d, Extended Data Fig. 1k). This finding was unrelated to changes in volume homeostasis, as estimated plasma volume and haematocrit were similar among the cohorts (Extended Data Fig. 1l, m).

We first determined whether the protective effect was due to expression of ANP by using ANP-releasing osmotic pumps implanted subcutaneously into mice before *C. novyi*-NT treatment. ANP delivered by pumps proved efficacious, with 75% of the mice surviving (Fig. 1a, upper panel), 77% of which exhibited complete tumour eradications. The other 23% of mice showed a robust but not curative response (Fig. 1a, bottom panel). Similar to ANP-*C. novyi*-NT, systemically delivered ANP also markedly reduced pro-inflammatory cytokines, catecholamines and tissue injury (Fig. 1 c, d, Extended Data Fig. 1h–k). Lastly, the effects of ANP were confirmed in another tumour model using subcutaneous implants of the glioblastoma cell line GL-261 in C56Bl/6 mice (Extended Data Fig. 2a).

[1]Ludwig Center and the Howard Hughes Medical Institute at the Johns Hopkins Kimmel Cancer Center, Baltimore, MD, USA. [2]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [3]Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [4]Department of Pathology, University of Washington, Seattle, WA, USA. [5]H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. [6]Johns Hopkins University School of Medicine, Baltimore, MD, USA. [7]These authors contributed equally: Verena Staedtke, Ren-Yuan Bai. *e-mail: vstaedt1@jhmi.edu; rbai1@jhmi.edu; vogelbe@jhmi.edu; sbzhou@jhmi.edu

© 2018 Springer Nature Limited. All rights reserved.

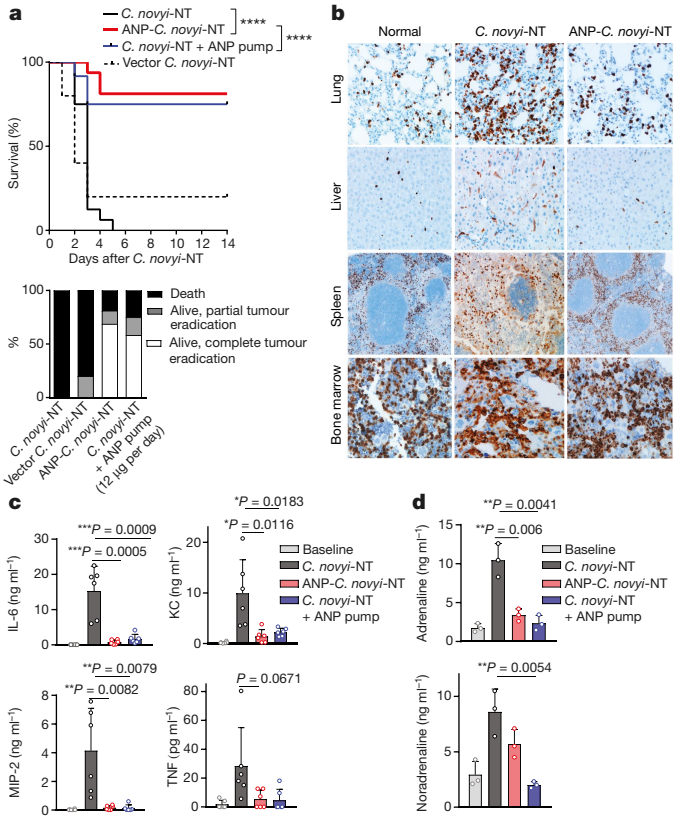


Fig. 1 | ANP reduces mortality. a, Kaplan–Meier curve (top panel) and therapeutic response (bottom panel) of ANP-*C. novyi*-NT ($n = 16$) compared to *C. novyi*-NT ($n = 16$), *C. novyi*-NT with ANP via osmotic pump ($n = 12$) and vector *C. novyi*-NT control ($n = 5$). Statistical survival differences were evaluated by two-sided log-rank test. $^{****}P \leq 0.0001$. **b**, Representative anti-CD11b-antibody-stained sections from the lungs, liver, spleen and bone marrow of mice treated with *ANP-C. novyi*-NT ($n = 3$) and *C. novyi*-NT ($n = 3$) compared to normal controls ($n = 2$). **c**, Plasma levels of indicated cytokines ($n = 6$ independent samples per group) 36 h after spore injection. **d**, Corresponding plasma levels of adrenaline and noradrenaline 36 h after *C. novyi*-NT, ANP-*C. novyi*-NT and *C. novyi*-NT plus ANP pump compared to normal controls ($n = 3$ per group). Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**c**, **d**).

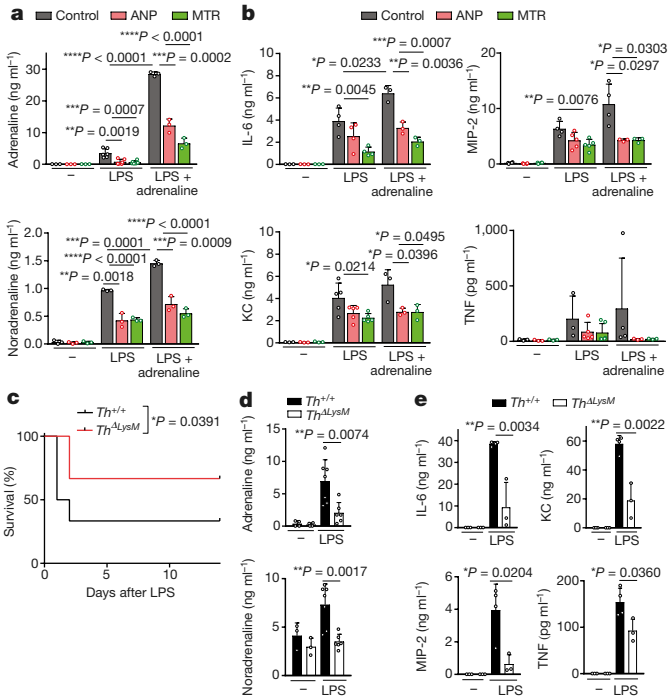


Fig. 2 | Catecholamine production in myeloid cells is essential for cytokine release. a, Peritoneal macrophages were pre-incubated with ANP or MTR for 10 min and then stimulated with LPS (50 μg ml^{-1}) or a combination of LPS and adrenaline (15 ng ml^{-1}) in vitro. Shown are the levels of adrenaline (left to right, $n = 3, 3, 3, 6, 6, 6, 3, 3, 3$ per column) and noradrenaline ($n = 3$) in the supernatant after 24 h. **b**, Corresponding cytokines from macrophage culture supernatants: IL-6 ($n = 3, 3, 3, 4, 4, 4, 3, 3, 3$), MIP-2 ($n = 4, 4, 4, 4, 5, 5, 4, 3, 3$), KC ($n = 3, 3, 3, 5, 5, 5, 3, 3, 3$) and TNF ($n = 3, 3, 3, 3, 5, 6, 4, 3, 3$). **c**, Survival of $Th^{+/+}$ and $Th^{\Delta LysM}$ mice treated with LPS and analysed with two-sided log-rank test ($n = 12$; 6 male, 6 female). **d**, **e**, Plasma levels of adrenaline ($n = 4, 4, 7, 6$) and noradrenaline ($n = 3, 3, 7, 6$) (**d**) and indicated cytokines ($n = 3, 3, 4, 3$) (**e**) at baseline and 24 h after LPS treatment in $Th^{+/+}$ or $Th^{\Delta LysM}$ mice. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**a**, **b**, **d**, **e**).

We then investigated the mechanism underlying the protective effects of ANP. Previous studies link the anti-inflammatory properties of ANP to inhibition of phosphorylation of inhibitory κ B protein (IκB)[9]. Treatment with BMS-345541, a highly selective IκB kinase inhibitor[11], did not improve survival in mice treated with *C. novyi*-NT (Extended Data Fig. 2b), suggesting that ANP inhibits inflammation resulting from *C. novyi*-NT through mechanisms in addition to the NF-κB pathway. This prompted us to investigate the relationship between catecholamines and ANP.

Macrophages, which are major sources of inflammatory cytokines, secrete and respond to catecholamines through adrenergic receptors when exposed to inflammatory stimuli such as bacteria[12,13]. This in turn leads to increases in cytokine production, as shown in models of lung injury and experimental autoimmune encephalomyelitis[13,14]. Given the pleiotropic effects of catecholamines, we first determined which catecholamine contributed to the severity of inflammation injury by using subcutaneously implanted osmotic pumps that continuously released adrenaline, noradrenaline or dopamine into mice treated with LPS. Only mice with adrenaline pumps showed an exacerbated disease course, with increased mortality and higher levels of IL-6, TNF and KC compared to LPS-treated controls and mice treated with adrenaline only, in which an increase of cytokines was also observed (Extended Data Fig. 3a–d).

We next investigated the effect of ANP on catecholamine synthesis in stimulated mouse peritoneal macrophages. ANP inhibited the upregulated production of macrophageal catecholamines induced by LPS, which correlated with a reduction in levels of IL-6, TNF, MIP-2 and KC (Fig. 2a, b, Extended Data Fig. 3e). Notably, LPS in combination with adrenaline produced a markedly enhanced inflammatory response compared to that observed with each of the other agents, and this amplification was also inhibited by ANP (Fig. 2a, b, Extended Data Fig. 3e–g). Direct inhibition of catecholamine synthesis with α-methyltyrosine (metyrosine, MTR), which blocks the key target tyrosine hydroxylase (TH) and prevents the conversion of tyrosine to L-DOPA, greatly reduced levels of catecholamines produced by stimulated mouse macrophages (Fig. 2a, Extended Data Fig. 3e, f). Accordingly, cytokines released by macrophages were also diminished by MTR (Fig. 2b, Extended Data Fig. 3g). Comparable results were obtained with human U937-derived macrophages (Extended Data Fig. 4a, b).

To confirm that the production of catecholamines by macrophages drives the inflammatory response, we used peritoneal macrophages from mice with selective deletion of *Th* gene in LysM^{+} myeloid cells[15] (LysM^{cre}*Th*$^{fl/fl}$ or $Th^{\Delta LysM}$) resulting in significantly reduced TH expression levels (Extended Data Fig. 4c). Peritoneal macrophages with *Th* deleted showed reduced secretion of catecholamines and cytokines upon stimulation with LPS and adrenaline, which confirmed the role of autocrine catecholamine production in the amplification of the inflammatory cascade in macrophages (Extended Data Fig. 4d, e).

© 2018 Springer Nature Limited. All rights reserved.

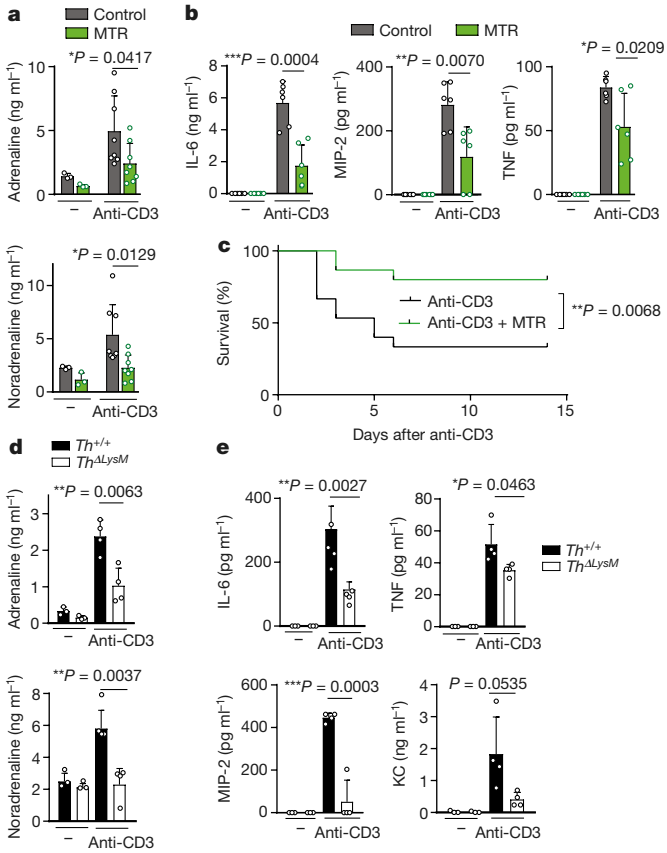


Fig. 3 | Inhibition of catecholamine synthesis reduces CRS after anti-CD3 treatment. a, **b**, Levels of adrenaline and noradrenaline (left to right, $n = 3, 3, 8, 8$ independent samples per column) (**a**) and of cytokines ($n = 6$ independent samples) (**b**) measured 24 h after anti-CD3 treatment, with or without MTR. **c**, Survival of BALB/c mice treated with anti-CD3, with or without MTR ($n = 15$ animals); analysed by two-sided log-rank test. **d**, **e**, Levels of adrenaline, noradrenaline ($n = 3, 3, 4, 4$) (**d**) and indicated cytokines ($n = 3, 3, 4, 4$) (**e**) measured 24 h after anti-CD3 treatment in $Th^{+/+}$ or $Th^{\Delta LysM}$ mice. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**a**, **b**, **d**, **e**).

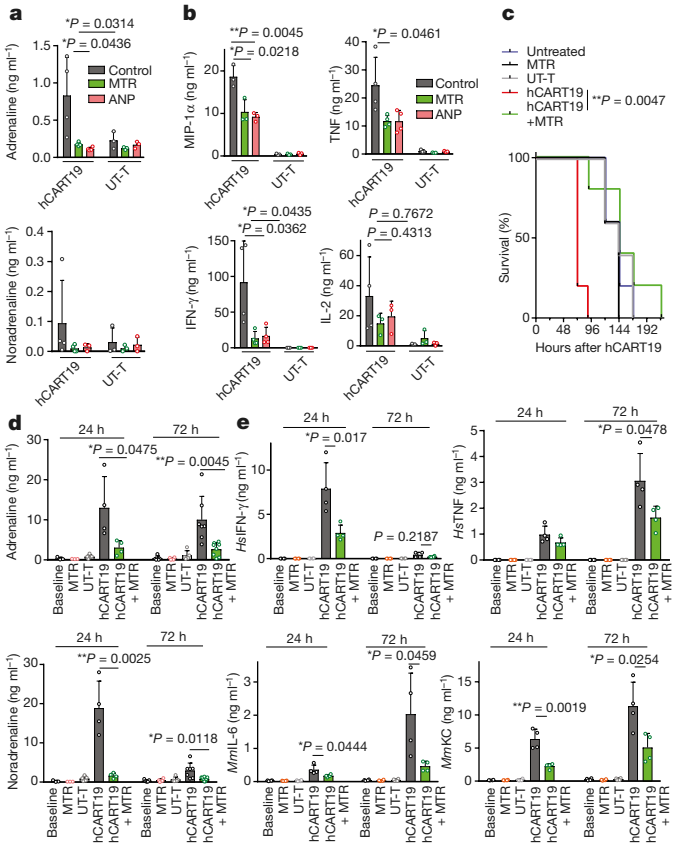


Fig. 4 | Inhibition of catecholamine synthesis reduces cytokine release induced by hCART19 in vitro and in vivo. a, **b**, Levels of adrenaline (left to right, $n = 4, 4, 4, 3, 3, 3$ per column) and noradrenaline ($n = 4, 4, 3, 3, 3, 3$) (**a**) and corresponding cytokines MIP-1α ($n = 3$), TNF ($n = 4, 4, 4, 3, 3, 3$), IFN-γ ($n = 4, 4, 4, 3, 3, 3$) and IL-2 ($n = 4, 3, 3, 3, 3, 3$) (**b**) in the supernatant 24 h after incubation of Raji cells with hCART19 or UT-T (ratio 1:5), with or without MTR or ANP. **c**, Survival of Raji-bearing NSGS mice with high tumour burden, treated with 1.5×10^7 hCART19, with or without MTR pre-treatment compared to UT-T, MTR and no treatment ($n = 5$ mice per group). Survival differences were evaluated by two-sided log-rank test. **d**, **e**, Levels of circulating adrenaline and noradrenaline ($n = 3, 3, 5, 4, 4, 5, 4, 5, 7, 8$) (**d**) and of indicated circulating mouse and human cytokines ($n = 4$ samples per group) (**e**), assessed 24 and 72 h after administration of hCART19 with or without MTR in comparison to controls. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**a**, **b**, **d**, **e**).

Notably, the impaired ability to produce catecholamines led to a significant reduction in LPS-induced mortality and cytokine release in $Th^{\Delta LysM}$ mice (Fig. 2c–e).

MTR was found to have similar effects in vivo. Around 75% of mice injected with LPS survived when pre-treated with MTR compared to only 10% of control mice (Extended Data Fig. 5a). The effect of MTR on survival, catecholamines and cytokines was dose-dependent and 24-h serial plasma sampling showed sustained catecholamine and cytokine suppression (Extended Data Fig. 5a–e). To determine the relevant receptor, we used the inhibitors prazosin, RX 821002, metoprolol and ICI 118551 to block α_1, α_2, β_1 and β_2-adrenergic receptors, respectively[13]. Only blockade of α_1-adrenergic receptors by prazosin was effective in LPS-treated mice, achieving results similar to those obtained with MTR (Extended Data Fig. 6a–c).

To confirm the generality of these findings, we treated mice with MTR before the induction of CRS by infection with *C. novyi*-NT. Of the mice pre-treated with MTR, 85% survived, whereas only 8% of control mice survived (Extended Data Fig. 7a). As predicted, levels of catecholamines and cytokines were substantially reduced in the cohort pre-treated with MTR (Extended Data Fig. 7b, c).

Genetically engineered Gram-negative bacteria are also used in experimental therapies for cancer[16] and it is known that sepsis resulting from infection with Gram-negative bacteria differs from that caused by infection with Gram-positive bacteria, such as *C. novyi*-NT[17]. We therefore evaluated the effect of MTR in the caecal ligation and puncture (CLP) model, in which enteric bacteria, including many Gram-negative species cause polymicrobial peritoneal sepsis. MTR also significantly reduced the mortality from peritoneal sepsis: 22% of the mice survived the acute phase, whereas all control animals died (Extended Data Fig. 7d). When MTR was used in combination with the β-lactam antibiotic imipenem, more than two thirds of the mice survived CLP, whereas more than 90% of mice treated with imipenem alone died (Extended Data Fig. 7d). This result highlights that death from overwhelming bacterial infections is caused by both bacteria and host reaction (that is, CRS). To confirm that the detrimental host response was diminished by pre-treatment with MTR, we documented the expected effects of MTR on circulating catecholamines and cytokines (Extended Data Fig. 7e, f).

CRS is also observed after the administration of non-bacterial biotherapeutics and particularly those that activate T cells. For example, targeting the CD3 molecules of T cells with antibodies (muromonab-CD3, also known as OKT3) can mitigate autoimmunity and allograft rejection but leads to activation of T cells and CRS[18]. Accordingly, we found that cytokine release induced by anti-mouse-CD3 antibody 145-2C11 was accompanied by an upsurge of catecholamines

© 2018 Springer Nature Limited. All rights reserved.

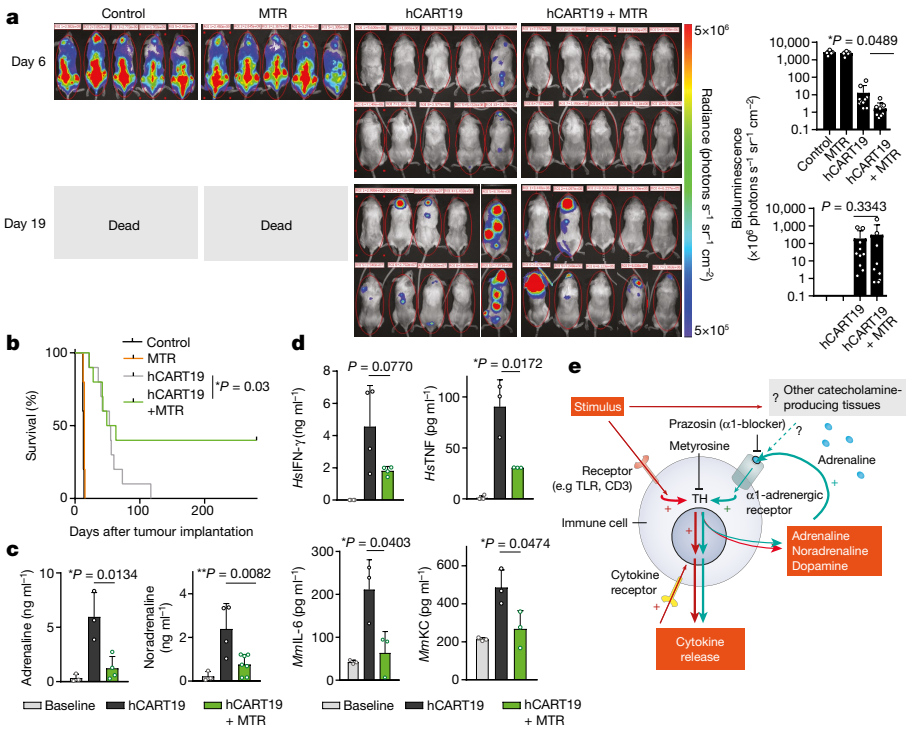


Fig. 5 | Inhibition of catecholamine synthesis with MTR does not impair the therapeutic response of hCART19. a, Serial bioluminescence imaging (BLI) of Raji-bearing NSGS mice (low tumour burden)at day 6 and 19 after treatment with 1.5×10^7 hCART19, with or without MTR ($n = 10$ mice per group) compared to control (UT-T), with or without MTR ($n = 5$ mice per group). BLI counts were used to quantify the tumour burden during the treatment course (right). Statistical differences were evaluated by one-tailed *t*-test. **b**, Corresponding Kaplan–Meier curve of Raji-bearing NSGS mice with low tumour burden, treated with 1.5×10^7 hCART19, with or without MTR pre-treatment ($n = 10$ mice per group) in comparison to control (UT-T), with or without MTR ($n = 5$ mice per group). Survival differences were analysed by weighted log-rank test (see Methods). **c**, **d**, Levels of plasma adrenaline ($n = 3, 3, 4$ per column) and noradrenaline ($n = 3, 4, 7$) (**c**) and *Hs*IFN-γ ($n = 4$), *Hs*TNF ($n = 4, 3, 3$), and mouse cytokines *Mm*IL-6 ($n = 3$) and *Mm*KC ($n = 3$) (**d**), assessed 72 h after hCART19 treatment. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test. **e**, Scheme showing how inhibition of the catecholamine pathway may reduce CRS. TLR, toll-like receptor.

in 5–6-month-old BALB/c mice (Fig. 3a–c, Extended Data Fig. 7g, h). Pre-treatment with MTR abrogated the increase in levels of catecholamines (Fig. 3a, Extended Data Fig. 7g) and several cytokines (IL-6, TNF, MIP-2, KC), whereas IL-2 and IFN-γ were unaffected (Fig. 3b, Extended Data Fig. 7h). Pre-treatment with MTR also protected against CRS-associated mortality (Fig. 3c). Similar results were observed in mice with myeloid-specific deletion of *Th*; these mice were protected from excessive catecholamine and cytokine release, indicating that myeloid-derived catecholamines are an essential mediator for CRS (Fig. 3d, e).

Genetically engineered T cells that express tumour-directed chimaeric antigen receptors (CARTs) often induce life-threatening CRS. Blockade of IL-6R and IL-1R has been used to suppress CRS in patients and experimental animals[19–21]. To investigate whether CARTs generate and release appreciable amounts of catecholamines during tumour cell killing, human Burkitt's lymphoma-derived $CD19^+$ Raji cells were incubated in vitro with human CD19-directed CARTs (denoted as hCART19 cells; CD19scFv-CD28-4-1BB-CD3ζ), as detailed in the Methods. hCART19–Raji cell interaction caused the release of catecholamines and cytokines (IL-2, TNF, IFN-γ, MIP-1α) and both MTR and ANP abated this reaction (Fig. 4a, b, Extended Data Fig. 8a). To demonstrate adrenaline-driven autocrine induction, we added adrenaline to co-cultured Raji and hCART19 cells and observed an amplified catecholamine and cytokine response (Extended Data Fig. 8a–c). This response was strongly inhibited by the protein synthesis inhibitor cycloheximide (CHX), indicating that de novo protein synthesis is required (Extended Data Fig. 8d, e).

To investigate the role of the catecholaminergic pathway on CART19-induced CRS in vivo, we injected Raji cells into sublethally irradiated[22,23] adult triple transgenic NSG-SGM3 (NSGS) mice. These mice express human myeloid supporting cytokines (IL3, GM-CSF, SCF) and can partially recapitulate CRS[22,24]. hCART19 cells were infused into Raji-bearing NSGS mice in two settings. In the first setting, hCART19 cells were infused at the half time of the median survival of untreated mice to establish a condition in which hCART19 cells would meet a high tumour burden and high risk for CRS (Fig. 4c–e, Extended Data Fig. 9a). We expected all these mice to die because hCART19 cells cannot rescue mice with high tumour burdens. In the second setting, hCART19 cells were infused at a third of the median survival time, when the mice had a relatively low tumour burden which allowed assessment of the anti-tumour response (Fig. 5a–d, Extended Data Fig. 9d). hCART19-treated mice with a high tumour burden died prematurely, with excessive levels of systemic catecholamines and cytokines at the time of death, including several human (*Hs*) T-cell-derived cytokines (*Hs*IL-2, *Hs*IFN-γ, and *Hs*TNF) and mouse (*Mm*) cytokines (*Mm*IL-6, *Mm*KC and *Mm*MIP-2) (Fig. 4c–e; Extended Data Fig. 9a–c). Pre-treatment with MTR significantly lowered the levels of circulating catecholamines and cytokines (*Hs*IFN-γ, *Hs*TNF, *Mm*IL-6, *Mm*KC and *Mm*MIP-2) but animals ultimately died, as expected, from progressive disease (Fig. 4c–e; Extended Data Fig. 9a–c). In mice with low tumour burdens, substantial anti-tumour effects of hCART19 cells were observed and more so in mice that received MTR (Fig. 5a, b; Extended Data Fig. 9d). Pre-treatment with MTR significantly lowered the increase of catecholamines, *Hs*TNF, *Mm*IL-6 and *Mm*KC in this model, and to a lesser degree *Hs*IFN-γ as well as *Hs*IL-2 (Fig. 5c, d; Extended Data Fig. 9e). We repeated these experiments with ANP, which yielded similar results (Extended Data Fig. 9f–h). Neither MTR nor ANP substantially interfered with hCART19 expansion (Extended Data Fig. 9i) or tumour clearance, and both were effective at preventing the cytokine release. Animals treated with the same amount of

© 2018 Springer Nature Limited. All rights reserved.

untransduced T cells (UT-T) did not show changes in survival, catecholamine or cytokine levels.

Because xenograft mouse models cannot fully predict the clinical behaviour of CART cells, we tested the effects of MTR and ANP in a syngeneic mouse model. C57Bl/6 mice engrafted with Eμ-ALL cells, a leukemia cell line derived from an Eμ-myc transgenic mouse to develop CD19-positive B cell acute lymphoblastic leukemia (B-ALL), were treated with CART19 cells directed against mouse CD19 (m1928z, mCART19), as detailed in Methods. Pre-treatment of mice with ANP or MTR did not affect the efficacy of the mCART19 cells in this model. However, the systemic release of catecholamines and cytokines was reduced by ANP and MTR, thereby confirming that these drugs may prevent the cytokine release while maintaining anti-tumour efficacy (Extended Data Fig. 10a–d).

A model illustrating the role of catecholamines in CRS is depicted in Fig. 5e. Our data, combined with previous studies[13,14,25], suggest that catecholamines enhance inflammatory injury resulting from bacterial and non-bacterial causes through a self-amplifying feed-forward loop in myeloid cells. Other catecholamine-producing cells, such as adrenal cells and T cells, in which stimulus-induced elevation of α_1- and α_2-adrenergic receptor levels have been reported[26], also probably participate in this feed-forward loop. Catecholamines secreted by such cells are synthesized by TH and act through α_1-adrenergic receptors expressed by immune cells. This circuit can be pharmacologically interrupted to modulate the inflammatory response. As MTR and prazosin are approved by the Food and Drug Administration for the treatment of hypertension, clinical translation of the findings is possible in clinical trials of agents whose application is restricted by CRS.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0774-y.

Received: 5 December 2017; Accepted: 22 October 2018;
Published online 12 December 2018.

1. van der Poll, T., van de Veerdonk, F. L., Scicluna, B. P. & Netea, M. G. The immunopathology of sepsis and potential therapeutic targets. *Nat. Rev. Immunol.* **17**, 407–420 (2017).
2. Hansel, T. T., Kropshofer, H., Singer, T., Mitchell, J. A. & George, A. J. The safety and side effects of monoclonal antibodies. *Nat. Rev. Drug Discov.* **9**, 325–338 (2010).
3. Rommelfanger, D. M. et al. The efficacy versus toxicity profile of combination virotherapy and TLR immunotherapy highlights the danger of administering TLR agonists to oncolytic virus-treated mice. *Mol. Ther.* **21**, 348–357 (2013).
4. Maude, S. L., Barrett, D., Teachey, D. T. & Grupp, S. A. Managing cytokine release syndrome associated with novel T cell-engaging therapies. *Cancer J.* **20**, 119–122 (2014).
5. Parker, B. S., Rautela, J. & Hertzog, P. J. Antitumour actions of interferons: implications for cancer therapy. *Nat. Rev. Cancer* **16**, 131–144 (2016).
6. Agrawal, N. et al. Bacteriolytic therapy can generate a potent immune response against experimental tumors. *Proc. Natl Acad. Sci. USA* **101**, 15172–15177 (2004).
7. Peters van Ton, A. M., Kox, M., Abdo, W. F. & Pickkers, P. Precision immunotherapy for sepsis. *Front. Immunol.* **9**, 1926 (2018).
8. Weber, G. F. et al. Interleukin-3 amplifies acute inflammation and is a potential therapeutic target in sepsis. *Science* **347**, 1260–1265 (2015).
9. Vollmar, A. M. The role of atrial natriuretic peptide in the immune system. *Peptides* **26**, 1086–1094 (2005).
10. Kuehne, S. A. & Minton, N. P. ClosTron-mediated engineering of *Clostridium*. *Bioengineered* **3**, 247–254 (2012).
11. Burke, J. R. et al. BMS-345541 is a highly selective inhibitor of I kappa B kinase that binds at an allosteric site of the enzyme and blocks NF-kappa B-dependent transcription in mice. *J. Biol. Chem.* **278**, 1450–1456 (2003).
12. Johnson, J. D. et al. Catecholamines mediate stress-induced increases in peripheral and central inflammatory cytokines. *Neuroscience* **135**, 1295–1307 (2005).
13. Flierl, M. A. et al. Phagocyte-derived catecholamines enhance acute inflammatory injury. *Nature* **449**, 721–725 (2007).
14. Shaked, I. et al. Transcription factor Nr4a1 couples sympathetic and inflammatory cues in CNS-recruited macrophages to limit neuroinflammation. *Nat. Immunol.* **16**, 1228–1234 (2015).
15. Clausen, B. E., Burkhardt, C., Reith, W., Renkawitz, R. & Förster, I. Conditional gene targeting in macrophages and granulocytes using LysMcre mice. *Transgenic Res.* **8**, 265–277 (1999).
16. Zheng, J. H. et al. Two-step enhanced cancer immunotherapy with engineered *Salmonella typhimurium* secreting heterologous flagellin. *Sci. Transl. Med.* **9**, eaak9537 (2017).
17. Surbatovic, M. et al. Cytokine profile in severe Gram-positive and Gram-negative abdominal sepsis. *Sci. Rep.* **5**, 11355 (2015).
18. Sevmis, S. et al. OKT3 treatment for steroid-resistant acute rejection in kidney transplantation. *Transplant. Proc.* **37**, 3016–3018 (2005).
19. Neelapu, S. S. et al. Chimeric antigen receptor T-cell therapy - assessment and management of toxicities. *Nat. Rev. Clin. Oncol.* (2017).
20. Giavridis, T. et al. CAR T cell-induced cytokine release syndrome is mediated by macrophages and abated by IL-1 blockade. *Nat. Med.* **24**, 731–738 (2018).
21. Norelli, M. et al. Monocyte-derived IL-1 and IL-6 are differentially required for cytokine-release syndrome and neurotoxicity due to CAR T cells. *Nat. Med.* **24**, 739–748 (2018).
22. Gill, S. et al. Preclinical targeting of human acute myeloid leukemia and myeloablation using chimeric antigen receptor-modified T cells. *Blood* **123**, 2343–2354 (2014).
23. Ninomiya, S. et al. Tumor indoleamine 2,3-dioxygenase (IDO) inhibits CD19-CAR T cells and is downregulated by lymphodepleting drugs. *Blood* **125**, 3905–3916 (2015).
24. Wunderlich, M. et al. A xenograft model of macrophage activation syndrome amenable to anti-CD33 and anti-IL-6R treatment. *JCI Insight* **1**, e88181 (2016).
25. Flierl, M. A. et al. Upregulation of phagocyte-derived catecholamines augments the acute inflammatory response. *PLoS One* **4**, e4414 (2009).
26. Bao, J. Y., Huang, Y., Wang, F., Peng, Y. P. & Qiu, Y. H. Expression of α-AR subtypes in T lymphocytes and role of the α-ARs in mediating modulation of T cell function. *Neuroimmunomodulation* **14**, 344–353 (2007).

Acknowledgements We thank N. Minton, C. Brayton, K. Kammers, D. Pardoll, Z. Li, E. Watson, C. Thoburn, S. Roy and N. Forbes-McBean for scientific and technical support. This work was supported by NINDS R25NS065729 (V.S.), NCI 1K08CA230179-01 (V.S.), Francis S. Collins Scholar Program (V.S.), DHART-SPORE IN4689861JHU (V.S.), 1R03CA178118-01A1 (R.-Y.B.), BVD (S.Z, R.-Y.B.), the Virginia and D.K. Ludwig Fund for Cancer Research (N.P., K.W.K., B.V., S.Z.), the BKI at Johns Hopkins, and CA062924 (S.Z.).

Reviewer information *Nature* thanks K. Tracey and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions V.S. and R.-Y.B. originated the concept, designed and performed experiments, analysed the data and wrote the manuscript. K.K. and P.B.R. provided scientific advice. M.D. and M.L.D. provided animal models and scientific advice. B.V. and S.Z. designed the study, interpreted the data and wrote the manuscript. G.J.R., K.W.K. and N.P. contributed to the manuscript. All authors approved the final manuscript.

Competing interests A patent application on CRS prevention listing V.S., R.-Y.B., G.J.R., K.W.K., N.P., S.Z. and B.V. as co-inventors has been provisionally filed by Johns Hopkins University. Under a licensing agreement between BVD Inc. and Johns Hopkins University, K.W.K., B.V. and S.Z. are entitled to a share of royalties managed by Johns Hopkins University. B.V., K.W.K. and NP are members of the Scientific Advisory Board of Sysmex and are founders of PapGene and Personal Genome Diagnostics. B.V. is also an advisor to Camden Partners. The terms of all these arrangements are managed by Johns Hopkins University in accordance with its conflict of interest policies. P.B.R. serves on Merck's board of directors.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0774-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0774-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to V.S., R.-Y.B., B.V. or S.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. All animal experiments were randomized; the investigators were blinded to the allocation, treatment and outcome assessment of experiments involving *C. novyi*-NT, CLP and anti-CD3.

Mice. All animal experiments were performed in accordance with protocols approved by the Johns Hopkins Animal Care and Use Committee (ACUC) and relevant animal use guidelines and ethical regulations were followed. For subcutaneous CT26 tumour implantation, LPS and CLP experiments, female C57Bl/6 and BALB/C mice of 6–8 weeks were purchased from Harlan Laboratories. For anti-mCD3 treatment, female BALB/C mice of 5–6 months old were purchased form Harlan laboratories. For the CART19 treatment, NSG-SGM3 (NSGS) mice (stock number 013062) were purchased from the Jackson Laboratory.

LysM^{cre}-conditional *Th*-knockout mice. LysM^{cre} mice were purchased from Jackson Laboratory (stock number 004781), in which a nuclear-localized Cre recombinase was inserted into the first coding exon of the lysozyme 2 (*Lyz2* or *LysM*) gene and expressed in the myeloid cell lineage (monocytes, mature macrophages and granulocytes)[15]. *Th* loxP/loxP ($Th^{fl/fl}$) mice were provided by M. Darvas at the University of Washington[27]. By crossing these two strains, LysM$^{cre}Th^{fl/fl}$ mice ($Th^{\Delta LysM}$) were produced as an experimental strain for LPS and anti-CD3 experiments and LysM$^{cre}Th^{+/+}$ mice ($Th^{+/+}$) were used as the Cre transgene control.

Chemicals and reagents. Anti-mCD3 (145-2C11) and anti-mIL6 receptor (15A7) antibodies were purchased from BioXcell. Anti-mTNFα antibody (R023) was purchased from Sino Biological and anti-mIL3 antibody (MP2-8F8) was purchased from BD Biosciences. α-methyl-D,L-p-tyrosine methyl ester hydrochloride (Santa Cruz Biotechnology, SC-219470) is a soluble from of α-methyl-tyrosine (metyrosine, MTR), which is converted to α-methyl-tyrosine in vivo[28], whereas the less soluble α-methyl-tyrosine was purchased from Sigma (120693). LPS from *Escherichia coli* 0111:B4 (L2630), (−)-adrenaline (E4250), dopamine (H8502), noradrenaline (A7256), prazosin (P7791), metoprolol (M5391) and human ANP (A1663) were purchased from Sigma. RX 821002 (1324) and ICI 118551 (0821) were purchased from Tocris.

Strain engineering of *C. novyi*-NT. The site-specific knock-in of human ANP in *C. novyi*-NT employed the TargeTron Gene Knockout System (Sigma), which is based on the retrohoming mechanism of group II introns[10]. The sequence of the human ANP cDNA was optimized for *Clostridium* codon usage as 5'-TCATTAAGAAGATCTTCATGTTTTGGAGGAAGAATGGATAGAATAGG AGCTCAATCAGGATTAGGATGTAATTCATTCAGATATTAA-3' coding for 28 AA (SLRRSSCFGGRMDRIGAQSGLGCNSFRY). The synthesized sequence was cloned into the shuttle vector pMTL8325. The construct included the *C. novyi* PLC signal peptide sequence under the control of the *C. novyi* flagellin promoter. Subsequently, the *MluI* fragment of the construct was subcloned into the vector pAK001 (pMTL8325-pJIR750ai Reverse-pFla-153 s-MCS-pThio-G1-ErmB) targeting the knock-in in the 153S site of *C. novyi*-NT genome. The *E.coli* CA434 strain containing the targeting construct was conjugated with *C. novyi*-NT and selected with polymyxin B/erythromycin (Sigma) under anaerobic condition. Colonies were selected and re-plated three times on non-selection plates and again on the erythromycin plate. Clones were tested first by PCR using EBS Universal and 153S-F primers. Positive clones were further tested by PCR with primers targeting the backbone of the vector to confirm the insert was integrated in *C. novyi* genome and with primers covering externally both sides of 153S to confirm the correct insertion. The propagation and sporulation of *C. novyi*-NT strains followed procedures described previously[29].

RNA extraction and quantitative PCR of *C. novyi*-NT strains. For quantitative reverse transcription with PCR (RT–PCR), RNA of germinated *C. novyi*-NT strains were extracted using RiboPure Bacterial RNA Purification Kit (Ambion) and transcribed with SuperScript IV RT Kit (Invitrogen) as described[29]. Real-time PCR was performed using Maxima SYBR Green/ROX qPCR Master Mix (Thermo Fisher), targeting on the NT01CX1854 gene specific for geminating *C. novyi*-NT[29].

ANP measurement and cGMP assay. ANP concentrations in the supernatant of ANP-*C. novyi*-NT culture and in mouse plasma were measured with an ELISA kit from Ray Biotech (EIAR-ANP-1) that recognizes both human and mouse ANP. ANP in the supernatant of ANP-*C. novyi*-NT culture were shown to have biological activity as described before[30]. Briefly, bacterial supernatants were applied to cultured bovine aortic endothelial cells (BAOEC, Cell Applications Inc.) for 3 min. cGMP concentrations were then measured in BAOEC lysates by the Direct cGMP ELISA Kit from Enzo following the manufacture's instruction.

Subcutaneous tumour models and *C. novyi*-NT therapy. The colon cancer cell line CT26 was injected subcutaneously into the right flank of 6–8-week-old female BALB/C mice as described previously[6]. Tumour sizes were measured with a caliper and calculated as (L × W × H)/2. When tumours reached 600–900 mm^3 after about two weeks, 12×10^6 spores of *C. novyi*-NT or ANP-*C. novyi*-NT at 3×10^6 per μl were injected intratumourally into four central parts of the tumour with a 32G Hamilton syringe needle. The bacteria typically germinated in the tumours within 24 h, turning them necrotic. Hydration of the mice was supported by daily subcutaneous injections of 500 μl saline. Human ANP (Sigma) was dissolved in saline, loaded in mini-osmotic pumps (ALZET) with a release rate of 12 μg per day and implanted subcutaneously in the back of mice 12 h before the spore injection. Pumps loaded with saline served as controls. MTR was dissolved in PBS and injected intraperitoneally at 60 mg kg^{-1} per day for three days before the *C. novyi* injection to deplete catecholamines in storage. Two hours after the spore injection, 60 mg kg^{-1} of MTR was injected intraperitoneally. For each of the next three days, intraperitoneal injections of MTR at 30 mg kg^{-1} were administered. Control groups were injected with PBS at the same time points.

Immunohistochemistry. Immunostaining for CD11b was performed on formalin-fixed, paraffin-embedded sections on a Ventana Discovery Ultra autostainer (Roche Diagnostics) by S. Roy of JHU Oncology Tissue Services. Briefly, following dewaxing and rehydration on board, epitope retrieval was performed using Ventana Ultra CC1 buffer (6414575001, Roche Diagnostics) at 96 °C for 64 min. Primary antibody, anti-CD11b (1:8000 dilution; catalogue number ab133357, Abcam) was applied at 36 °C for 40 min. Primary antibodies were detected using an anti-rabbit HQ detection system (7017936001 and 7017812001, Roche Diagnostics) followed by Chromomap DAB IHC detection kit (5266645001, Roche Diagnostics), counterstaining with Mayer's haematoxylin, rehydration and mounting.

In vitro macrophage experiments. Isolation of elicited macrophages from mouse peritoneum followed previously described procedures with minor modifications[31]. Four days before collection, 1 ml of 3% Brewer's thioglycollate medium (BD) was injected intraperitoneally in female 2–3-month-old BALB/c mice or 4–6-week-old conditional TH-knockout mice. Mice were killed by cervical dislocation and the skin of the belly was cut open without penetrating the muscle layer. Using a syringe with a 25G needle, 5 ml of cold PBS containing 5 mM EDTA was injected carefully into the peritoneal cavity. After massaging gently for 1–2 min, a 1-ml syringe without needle was used to extract the peritoneal contents containing residential macrophages. Cells were centrifuged at 400*g* for 10 min at 4 °C, re-suspended in DMEM/F12 medium supplemented with 1% FBS and antibiotics and distributed in 48-well plates at a concentration of 0.5×10^6 cells/well. After incubation at 37 °C for 2 h, cells were rinsed three times with 0.5 ml medium and then 250 μl of medium was added to each well. Ten minutes before the addition of LPS or adrenaline, MTR at 2 mM or ANP at 5 μg ml^{-1} was added to the cells. For stimulation, the cells were incubated for 24 h with LPS at 50 μg ml^{-1}. An initial solution of 3 mg ml^{-1} (−)-adrenaline was made with 0.1N HCl and subsequently diluted with PBS. To stimulate macrophages, they were exposed to adrenaline at 15 ng ml^{-1} for 24 h at 37 °C. After the incubation, supernatants were collected from the wells and mixed with 5 mM EDTA and 4 mM sodium metabisulphite for preservation of catecholamines and stored at −80 °C. Control experiments showed that all detectable adrenaline was degraded after incubation in medium for 24 h at 37 °C. Thus, any adrenaline identified in the medium must have been secreted by cells in the last 24 h before collecting the medium.

Human U937 cells were cultured in RPMI 1640 medium with 5% FBS and antibiotics, and were differentiated to M1 macrophage-like cells by incubating with 20 nM phorbol 12-myristate 13-acetate (PMA, Sigma) for 24 h and further culturing in RPMI 1640 medium with 5% FBS and antibiotics for another 72 h. The experiments with U937 were set up in the same way as described above with peritoneal macrophages. Ten minutes before the addition of LPS or adrenaline, MTR at 2 mM or ANP at 5 μg ml^{-1} was added to the cells. Cells were incubated for 24 h with LPS at 1 μg ml^{-1}.

LPS experiments in mice. LPS from *Escherichia coli* 0111:B4 was formulated as a 10 mg ml^{-1} solution in water and stored in −80 °C. In BALB/C mice, LPS was injected intraperitoneally at a lethal dose of 3.5 mg kg^{-1}. This lethal dose was found to cause 70–90% death rate and be optimal for demonstrating the protective effects of ANP and MTR. In experiments with catecholamine pumps that were implanted a day before, a sublethal dose of LPS with 15–35% death rate was optimized in BALB/C mice. In $Th^{+/+}$ and $Th^{\Delta LysM}$ mice with C57Bl/6 background, a lethal dose of LPS was optimized at 5 mg kg^{-1}. Human ANP (Sigma) was dissolved in saline, loaded in mini-osmotic pumps (ALZET) with a release rate of 12 μg per day and implanted subcutaneously in the back of mice 12 h before the LPS injection. Mice implanted with pumps loaded with saline served as controls. MTR was freshly dissolved in PBS and injected intraperitoneally at the indicated doses for three days before the LPS treatment. One hour before the LPS injection, MTR was injected into the lower abdomen contralateral to the side of LPS injection. The control groups were injected with PBS. For the following 3 days, MTR was injected intraperitoneally at reduced indicated doses. Hydration of mice was supported by daily subcutaneous injection of 0.5 ml saline.

CLP experiments. CLP was performed as described previously[32]. Briefly, 6–8-week-old female C57Bl/6 mice were anesthetized and following abdominal incision, the caecum was ligated at about a quarter of the distance from the luminal

© 2018 Springer Nature Limited. All rights reserved.

entry to its tip. The ligated caecum was punctured through and through with a 22G needle at one half and three quarters of the distance from the luminal entry to its tip. A small amount of the caecal content was gently pushed out of the four openings into the peritoneum. Subsequently, the abdominal muscles were sutured and the skin was closed with two staples. Immediately after this, 500 ml of saline was injected subcutaneously to the mice. For the groups treated with antibiotics, imipenem (Sigma) was injected subcutaneously at 25 mg kg^{-1} starting from 20 h after CLP, with a schedule of twice a day on day one and once a day thereafter for 10 days. MTR was freshly dissolved in PBS and injected intraperitoneally at 60 mg kg^{-1} per day for three days before the CLP. Twenty minutes before the CLP, MTR was injected at 60 mg kg^{-1} intraperitoneally into the right side. The control groups were injected with PBS. For the following 4 days, MTR was injected at 30 mg kg^{-1} per day intraperitoneally into the right side. Hydration of mice was supported by daily subcutaneous injection of 0.5 ml saline.

Anti-CD3 treatment. For survival experiments, 5–6-month-old female BALB/c mice were used because we observed that young mice treated with anti-CD3 antibodies underwent severe weight loss but did not consistently die, even at very high doses of the anti-CD3 antibody. MTR was freshly dissolved in PBS and injected intraperitoneally at 60 mg kg^{-1} per day for three days before injection of anti-CD3 antibodies. Various doses of anti-CD3 antibody were tested, and it was found that 125 μg per mouse resulted in the death of about half the mice; this was the dose chosen for further experiments. Thirty minutes before the intraperitoneal injection of the anti-mouse CD3 antibody (BioXcell, 145-2C11), MTR was intraperitoneally injected at 60 mg kg^{-1} into the contralateral side. A single additional dose of 30 mg kg^{-1} MTR was injected intraperitoneally on the following day. Control groups were injected with PBS at the same times. For experiments with conditional TH-knockout mice, 4–6-week-old $LysM^{cre}Th^{fl/fl}$ ($Th^{\Delta LysM}$) mice with C57Bl/6 background were used and $LysM^{cre}Th^{+/+}$ mice of the same age were used as control. In these experiments, 200 μg per mouse anti-mouse CD3 antibody was injected intraperitoneally.

Human anti-CD19 CART (hCART19) cells and untransduced T cells (UT-T). Human CD19scFv-CD28-4-1BB-CD3ζ CAR-T cells (PM-CAR1003) were purchased from Promab Biotechnologies and stored in liquid nitrogen upon delivery. The CAR construct includes a scFv derived fromFMC63 anti-CD19 antibody, a hinge region and a transmembrane domain of CD28 in a third-generation CAR cassette. Generation of CAR-encoding lentivirus, isolation, expansion and transduction of human T cells followed the procedures published before by the manufacturer[33]. Cells were proliferated for two weeks in medium containing 300 IU ml^{-1} of human IL-2 by the manufacturer[33]. CART cells were used freshly upon defrosting or maintained less than 7 days in the CART medium consisting of AIM-V medium (GIBCO) supplemented with 5% FBS (Sigma) and penicillin-streptomycin (GIBCO), with the addition of 300 IU ml^{-1} of human IL-2 (Peprotech).

UT-T were purchased from ASTARTE Biologics (1017-3708OC17, $CD3^+$) and were used freshly upon defrosting or maintained less than 7 days in CART medium.

In vitro assays of hCART19 cells. Raji, a human Burkitt's lymphoma cell line, was purchased from Sigma. In a 48-well plate, Raji cells were plated at 1×10^5 per well and hCART19 cells or UT-T cells were plated at 5×10^5 per well in 275 μl of medium. A solution of 3 mg ml^{-1} (−)-adrenaline was made in 0.1N HCl and subsequently diluted in PBS for use at a final concentration of 15 ng ml^{-1}. Five minutes before the Raji and CART cells with or without adrenaline were mixed, MTR at 2 mM or human ANP at 5 μg $ml^{-1}$was added and then the cells were incubated for 24 h at 37 °C. Control experiments showed that all detectable adrenaline was degraded after incubation in medium for 24 h at 37 °C. Thus, any adrenaline identified in the medium must have been secreted by cells in the last 24 h before collecting the medium. Cycloheximide (CHX, Sigma) was added at 10 μg ml^{-1} to Raji and CART cells 30 min before they were mixed. After incubation, the cells were pelleted by centrifugation at 700*g* and 4 °C for 5 min and the supernatants were collected and mixed with 5 mM EDTA and 4 mM sodium metabisulphite for preservation of catecholamines, then stored at −80 °C until analysis.

Treatment of Raji tumour-bearing mice with hCART19 cells. We purchased 6–8-week-old female NSG-SGM3 (NSGS) mice (NOD.Cg-$Prkdc^{scid}$ $Il2rg^{tm1Wjl}$Tg (CMV-IL3, CSF2, KITLG) 1Eav/MloySzJ, stock number 013062) from the Jackson Laboratory. Raji cells were transfected with a luciferase construct via lentivirus to create Raji–luc cells. NSGS is a triple transgenic strain expressing human IL3, GM-CSF and SCF combining the features of the highly immunodeficient NOD scid gamma (NSG) mouse. One day before the injection of Raji cells, mice were irradiated at a dose of 2 Gy in a CIXD Xstahl device. In high tumour burden experiments in Fig. 4, 10^6 Raji–luc cells were injected intravenously through the tail vein. Six days later, tumour loads were assessed using a Xenogen instrument and 15×10^6 hCART19 cells or UT-T were injected intravenously. In low tumour burden experiments in Figs. 5, 2×10^5 Raji–luc cells were injected intravenously through the tail vein. Four days later, tumour loads were assessed using a Xenogen instrument and 15×10^6 hCART19 cells were injected intravenously. MTR was injected intraperitoneally at 60 mg kg^{-1} per day for 3 days before the hCART19 injection. On the day of CART19 injection, a fourth dose of 60 mg kg^{-1} was given intraperitoneally and the mice were subsequently injected four more times at daily intervals at 30 mg kg^{-1}.

Mouse anti-CD19 CART cells (mCART19) and untransduced T cells (UT-T). Mouse CD19scFv-CD28-CD3ζ CAR (m1928z) construct with GFP in SFG retroviral vector was described before and provided by M. Davila at the Moffitt Cancer Center[34]. The isolation, activation and transduction of mouse T cells followed the procedure described before[34,35]. Briefly, the spleens were collected from female C57Bl/6 mice and T cells were enriched from splenocytes by passage over a nylon wool column (Polysciences). Mouse T cells were then activated with CD3/CD28 Dynabeads (Thermo Fisher) following the manufacturer's instructions and cultured in the presence of human IL-2 at 30 IU ml^{-1} (R&D Systems). Retrovirus was produced by transfecting Phoenix-Eco packaging cells (ATCC) and spinoculations were done twice with retroviral supernatant. mCART19 cells were expanded for 10–14 days as described[35]. UT-T were produced following the same procedure without viral transduction.

Treating B cell acute lymphoblastic leukaemia (B-ALL) with mCART19 in immunocompetent mice. The Eμ-ALL cell line was derived from a lymphoid malignancy in an Eμ-myc transgenic mouse and upon intravenous injection, can develop B-ALL in C57Bl/6 mice[34]. The Eμ-ALL cells were provided by M. Davila and co-cultured with feeder NIH-3T3 cells that were irradiated at 60 Gy, in RPMI 1640 medium supplemented with 10% FBS, 0.05 mM 2-mercaptoethanol and antibiotics. Eμ-ALL cells were transfected with luciferase via lentivirus. 2×10^6 Eμ-ALL cells were intravenously injected in female 6–8-week-old C57Bl/6 mice through the tail vein and after 6 days, mice were intraperitoneally injected with cyclophosphamide (CPA) at 100 mg kg^{-1} for pre-conditioning as described before[34]. One day after CPA treatment, 10×10^6 mCART19 cells were intravenously injected in the mice. MTR was injected intraperitoneally at 40 mg kg^{-1} per day for 3 days before the mCART19 injection. On the day of mCART19 injection, a fourth dose of 40 mg kg^{-1} was given intraperitoneally and the mice were subsequently injected four more times at daily intervals at 30 mg kg^{-1}. One day before mCART19 injection, mini-osmotic pumps (ALZET) loaded with human ANP with a release rate of 12 μg per day were implanted subcutaneously in the back of mice. Tumour load was monitored by Xenogen before and after mCART19 injection.

Measurement of catecholamines and cytokines in mouse plasma. Blood samples were collected into tubes containing 5 mM EDTA and 4 mM sodium metabisulphite after puncturing the facial vein or (terminally) by cardiac puncture. Subsequently, the samples were centrifuged and the plasmas were stored at −80 °C before analysis. Catecholamines (dopamine, noradrenaline and adrenaline) were measured using the 3-CAT Research ELISA kit from Labor Diagnostika Nord GmbH/Rocky Mountain Diagnostics. Cytokines were measured using Luminex assays based on Millipore Mouse and Human Cytokine/Chemokine panels or ELISA kits for mouse or human IL-6, TNF, MIP-1α, KC, MIP-2 and IL-2 (R&D Systems) per manufacturer's instructions.

Statistical analysis. Statistical analysis was performed using GraphPad Prism 7 and R version 3.5.1. Statistical tests performed by Graphpad Prism included the two-tailed unpaired two-sample *t*-test; one-tailed unpaired two-sample *t*-test; the log-rank Mantel–Cox test; the Gehan–Breslow–Wilcoxon test; the weighted log-rank test using Fleming–Harrington weights was performed in R. The respective statistical test used for each figure is noted in the corresponding figure legends and significant statistical differences are noted as $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$, $**** P \leq 0.0001$.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

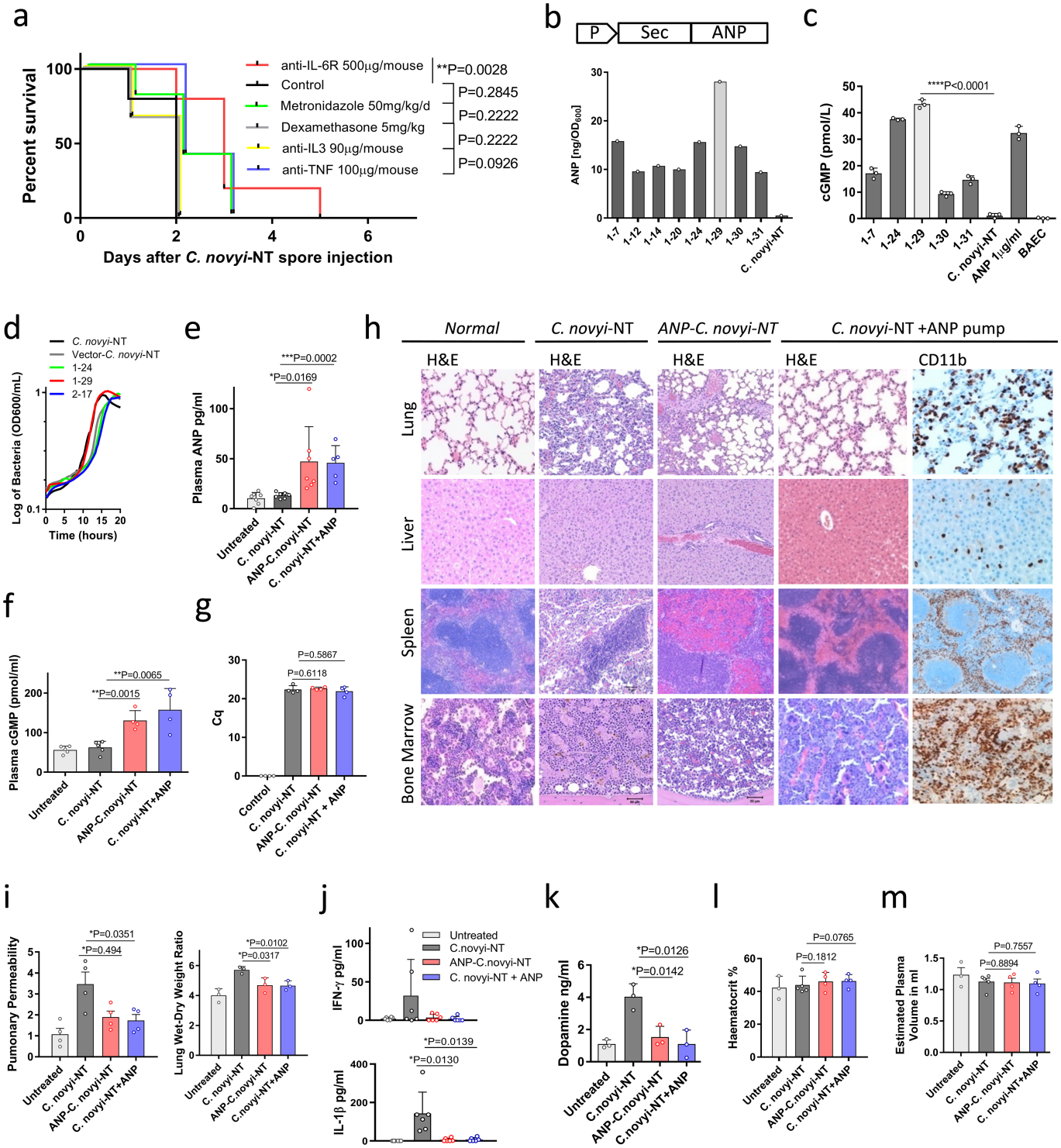
Source Data are provided for Figs. 1a, 2c, 3c, 4c, 5b and Extended Data Figs. 1a, 2a, b, 3a, b, 5a, 6a, 7a, 7d, 9h and 10d. The remaining datasets generated during this study are available from the corresponding authors on reasonable request. Unique materials such as the *C. novyi* strains are available on request to the corresponding authors. The transgenic mouse models and mouse CART constructs used in the study are made available through the original publishing authors.

27. Jackson, C. R. et al. Retinal dopamine mediates multiple dimensions of light-adapted vision. *J. Neurosci.* **32**, 9359–9368 (2012).
28. Corrodi, H. & Hanson, L. C. Central effects of an inhibitor of tyrosine hydroxylation. *Psychopharmacologia* **10**, 116–125 (1966).
29. Bettegowda, C. et al. The genome and transcriptomes of the anti-tumor agent *Clostridium novyi*-NT. *Nat. Biotechnol.* **24**, 1573–1580 (2006).
30. Lofton, C. E., Newman, W. H. & Currie, M. G. Atrial natriuretic peptide regulation of endothelial permeability is mediated by cGMP. *Biochem. Biophys. Res. Commun.* **172**, 793–799 (1990).
31. Zhang, X., Goncalves, R. & Mosser, D. M. The isolation and characterization of murine macrophages. *Curr. Protoc. Immunol.* https://doi.org/10.1002/0471142735.im1401s83 (2008).

© 2018 Springer Nature Limited. All rights reserved.

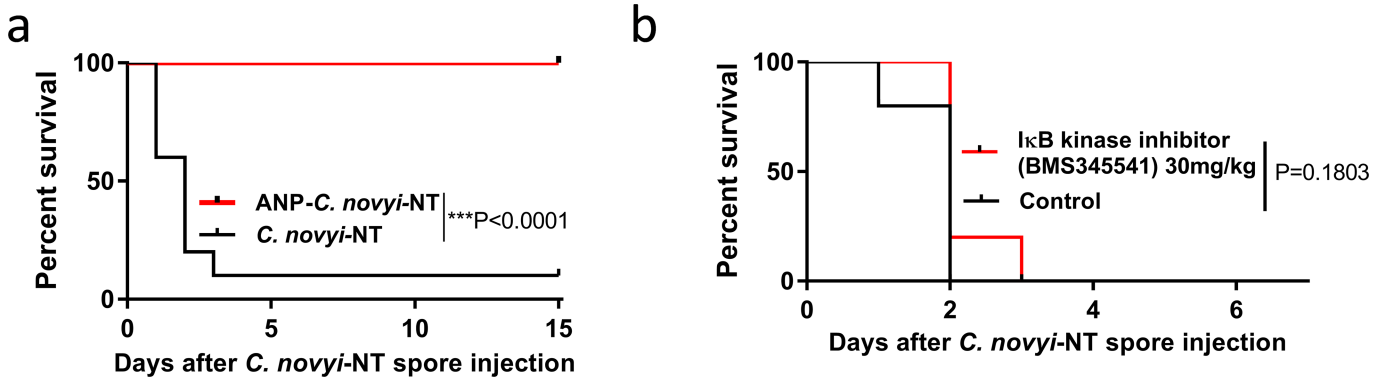
32. Rittirsch, D., Flierl, M. A. & Ward, P. A. Harmful molecular mechanisms in sepsis. *Nat. Rev. Immunol*. **8**, 776–787 (2008).
33. Berahovich, R. et al. FLAG-tagged CD19-specific CAR-T cells eliminate CD19-bearing solid tumor cells in vitro and in vivo. *Front. Biosci*. **22**, 1644–1654 (2017).
34. Davila, M. L., Kloss, C. C., Gunset, G. & Sadelain, M. CD19 CAR-targeted T cells induce long-term remission and B cell aplasia in an immunocompetent mouse model of B cell acute lymphoblastic leukemia. *PLoS One* **8**, e61338 (2013).
35. Lee, J., Sadelain, M. & Brentjens, R. Retroviral transduction of murine primary T lymphocytes. *Methods Mol. Biol*. **506**, 83–96 (2009).

© 2018 Springer Nature Limited. All rights reserved.



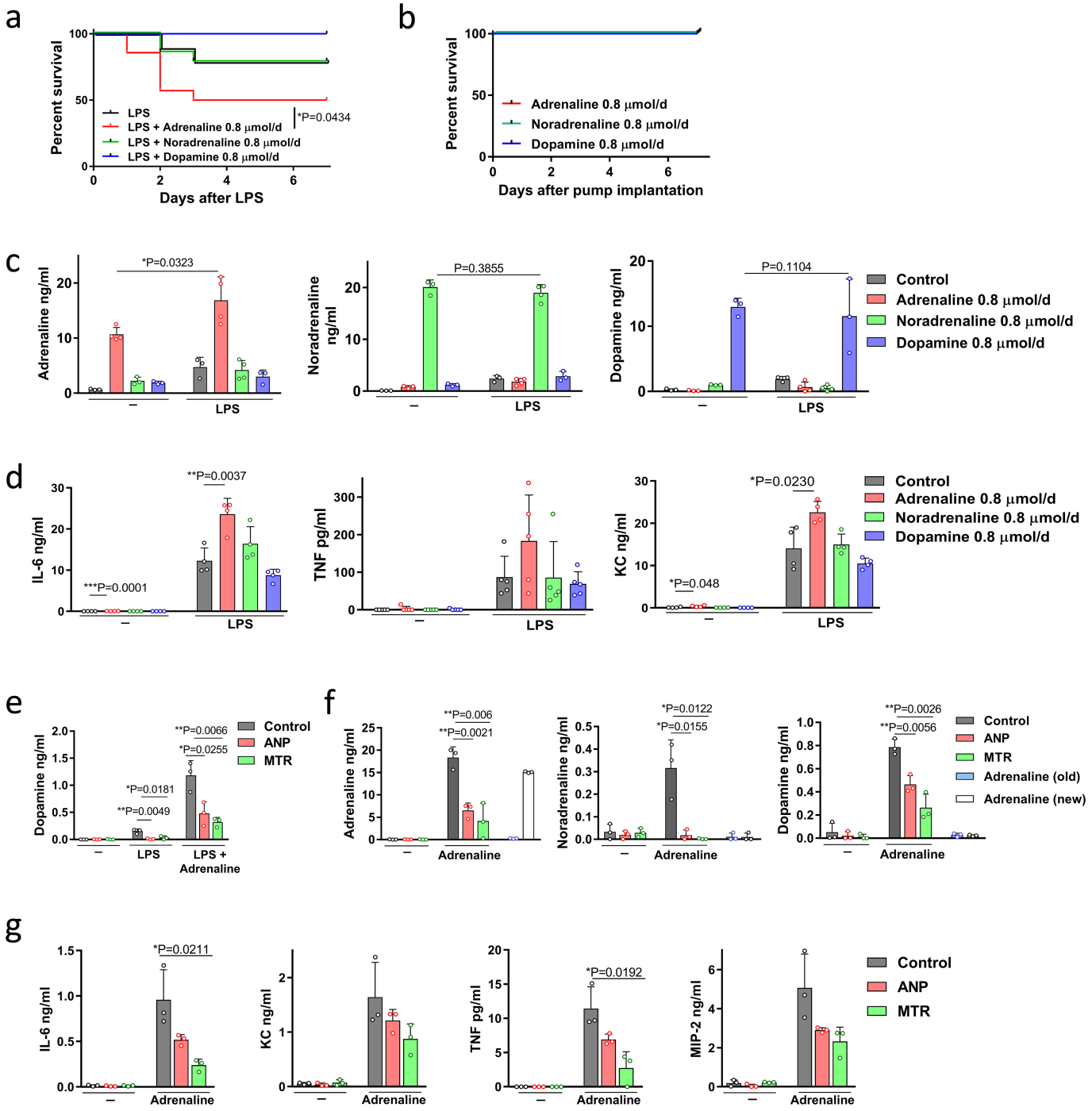
Extended Data Fig. 1 | In vitro and in vivo studies of ANP-*C. novyi*-NT. a, Kaplan–Meier curves of mice with large subcutaneous CT26 tumours (600–900 mm^3), treated with 12×10^6 *C. novyi*-NT spores and the indicated agents: anti-IL-6R ($n = 10$), metronidazole ($n = 5$), dexamethasone ($n = 6$), anti-IL3 ($n = 6$) and anti-TNF ($n = 5$) injected intratumourally, compared to controls ($n = 5$). Survival differences were analysed by two-sided log-rank test. **b** and **c**, Selected clones of ANP-*C. novyi*-NT were analysed for ANP secretion, shown as the average of a triplicate, (**b**) and for cGMP induction ($n = 3$) using bovine aortic endothelial cells (**c**). **d**, Growth pattern of several clones compared to the parental *C. novyi*-NT. The average of a triplicate is shown. **e**–**g**, Levels of plasma ANP (left to right, $n = 7, 8, 7, 5$ independent samples per column) (**e**), plasma cGMP ($n = 5, 5, 4, 4$ samples per column) (**f**) and germinated *C. novyi* strains in tumour tissue ($n = 4$ samples per column) based on quantification cycle (C_q) from RT–PCR of germination-specific NT01CX1854 gene (**g**), measured at 36 h after spore injection. **h**, Representative haematoxylin and eosin as well as anti-CD11b antibody stained sections from the lungs, liver, spleen and bone marrow of mice treated with *ANP-C. novyi*-NT ($n = 3$), *C. novyi*-NT ($n = 3$) and *C. novyi*-NT plus ANP ($n = 2$) compared to normal controls ($n = 2$). **i**–**m**, Pulmonary permeability ($n = 4$ mice per group), lung wet–dry ratio ($n = 3$ mice per group) (**i**) as well as levels of cytokines ($n = 6$ independent samples per column) (**j**), dopamine ($n = 3$ independent samples per column) (**k**), haematocrit ($n = 3, 5, 4, 4$ samples per column) (**l**) and calculated plasma volume ($n = 3, 5, 4, 4$ samples per column) (**m**) measured 36 h after spore treatment. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**c**, **e**–**g**, **i**–**m**). BAEC, bovine aortic endothelial cells.

© 2018 Springer Nature Limited. All rights reserved.



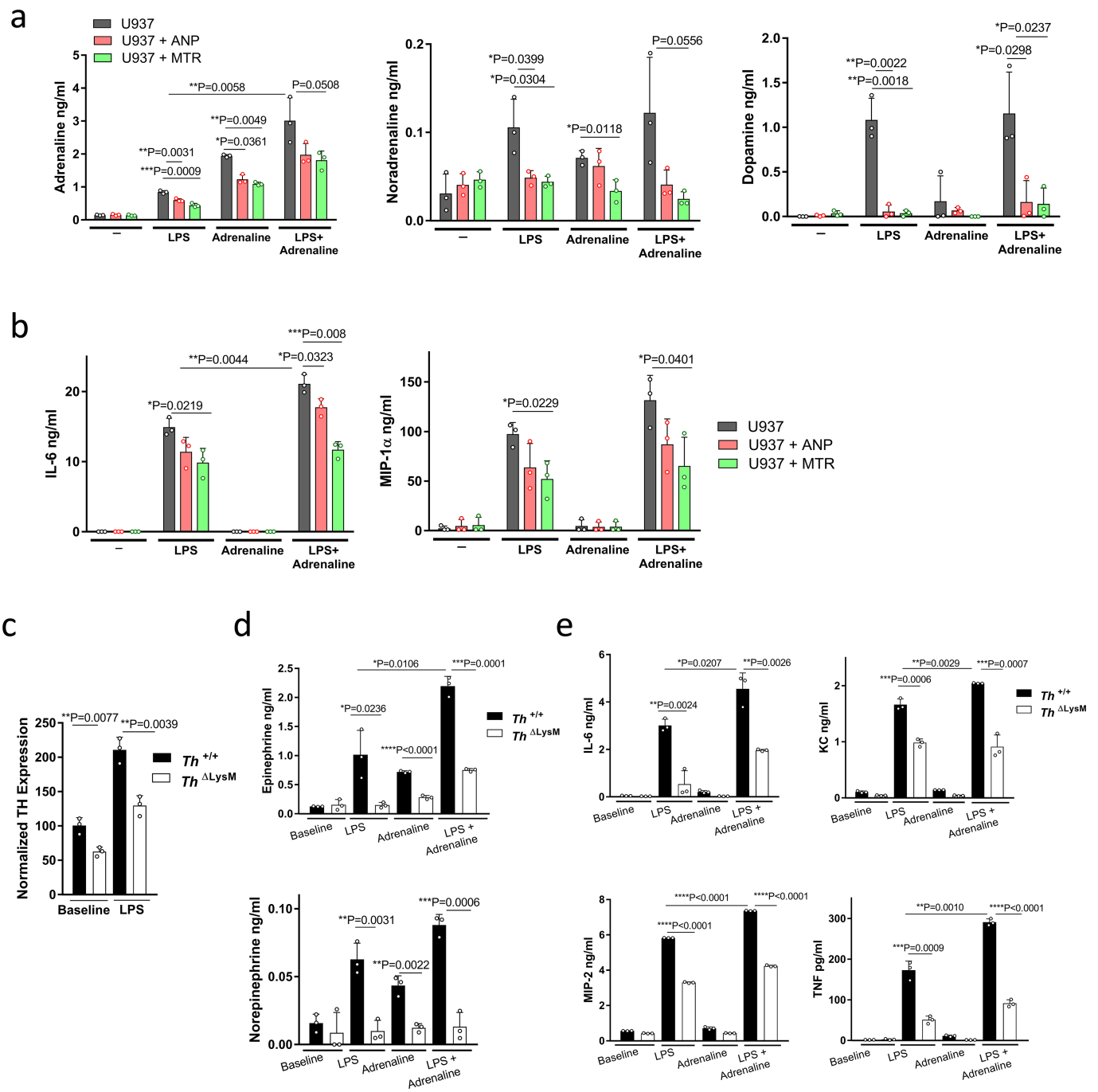
Extended Data Fig. 2 | Survival of mice treated with ANP and IκB kinase inhibitor BMS345541. a, Survival of mice with subcutaneously implanted GL-261 tumours, treated with 12×10^6 of ANP-*C. novyi*-NT spores ($n = 10$ animals per group). **b**, Survival of mice with CT26 tumours treated with *C. novyi*-NT and IκB kinase inhibitor BMS345541 ($n = 5$ mice per group). Survival differences were analysed by two-sided log-rank test (**b**, **c**).

© 2018 Springer Nature Limited. All rights reserved.



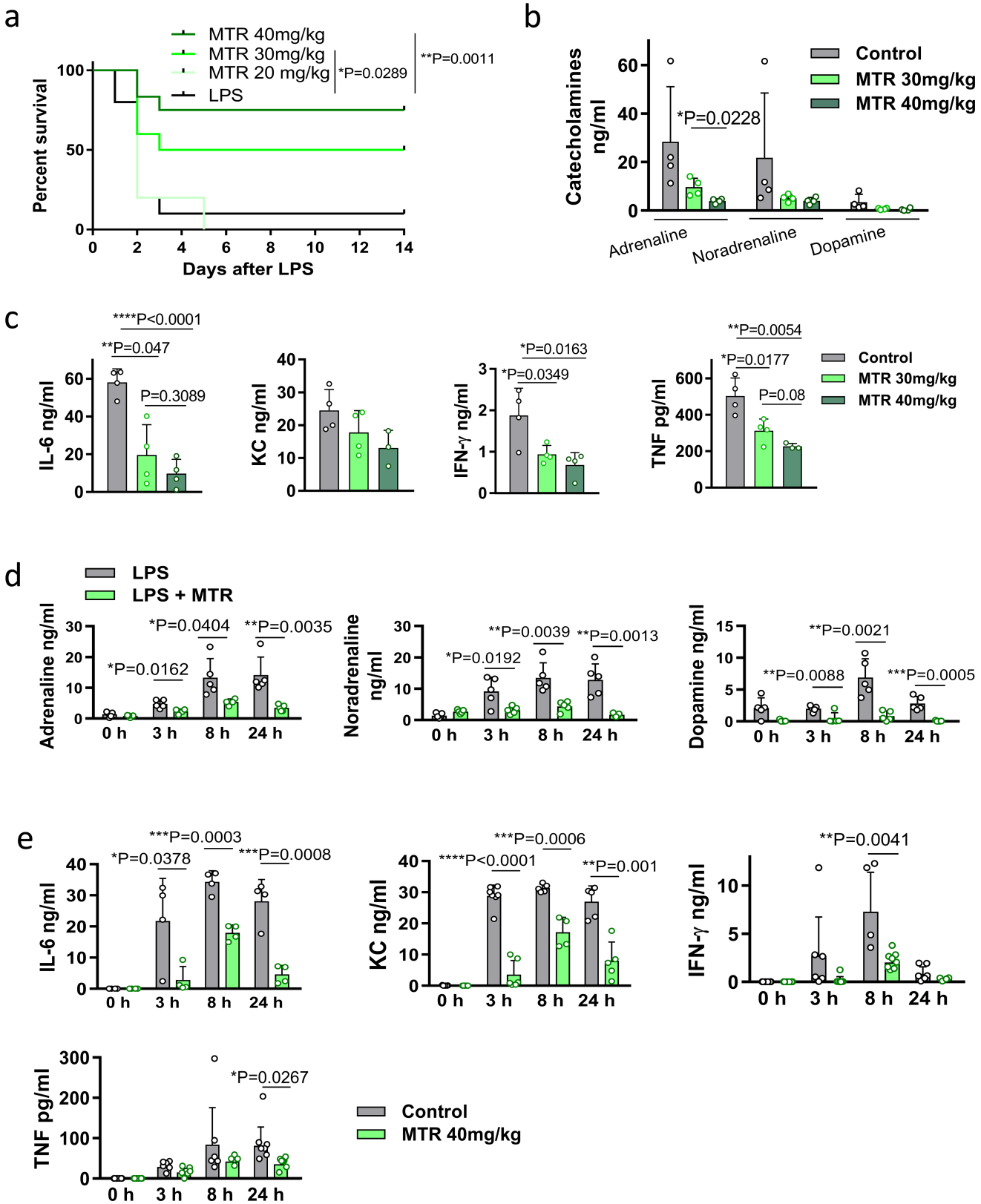
Extended Data Fig. 3 | Adrenaline enhances the inflammatory response. **a**, Survival of BALB/c mice implanted with the indicated catecholamine pump and stimulated with a sublethal dose of LPS ($n = 14$ mice per group) compared to LPS alone ($n = 19$ mice). Survival differences were analysed by Gehan–Breslow–Wilcoxon test. **b**, Survival of BALB/c mice with indicated catecholamine pump without LPS stimulation ($n = 5$ mice per group). **c**, **d**, 24 h plasma levels of adrenaline (left to right, $n = 3, 4, 3, 3, 3, 4, 4, 3$ per column), noradrenaline ($n = 3, 3, 3, 3, 3, 4, 4, 3$) and dopamine ($n = 3, 3, 3, 3, 3, 4, 4, 3$) (**c**) as well as levels of IL-6 ($n = 4$ per column), TNF ($n = 5$ per column) and KC ($n = 4$ per column) (**d**) in mice receiving the indicated treatments. **e**, Dopamine concentration in LPS- and adrenaline-treated peritoneal macrophages pre-incubated with ANP or MTR ($n = 3$ per column), measured after 24 h. **f**, **g**, Levels of catecholamines ($n = 3$ independent samples per column) (**f**) and several cytokines ($n = 3$ independent samples per column) (**g**) in adrenaline (15 ng ml^{-1})-treated peritoneal macrophages pre-incubated with ANP or MTR and measured after 24 h. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**c**–**g**).

© 2018 Springer Nature Limited. All rights reserved.



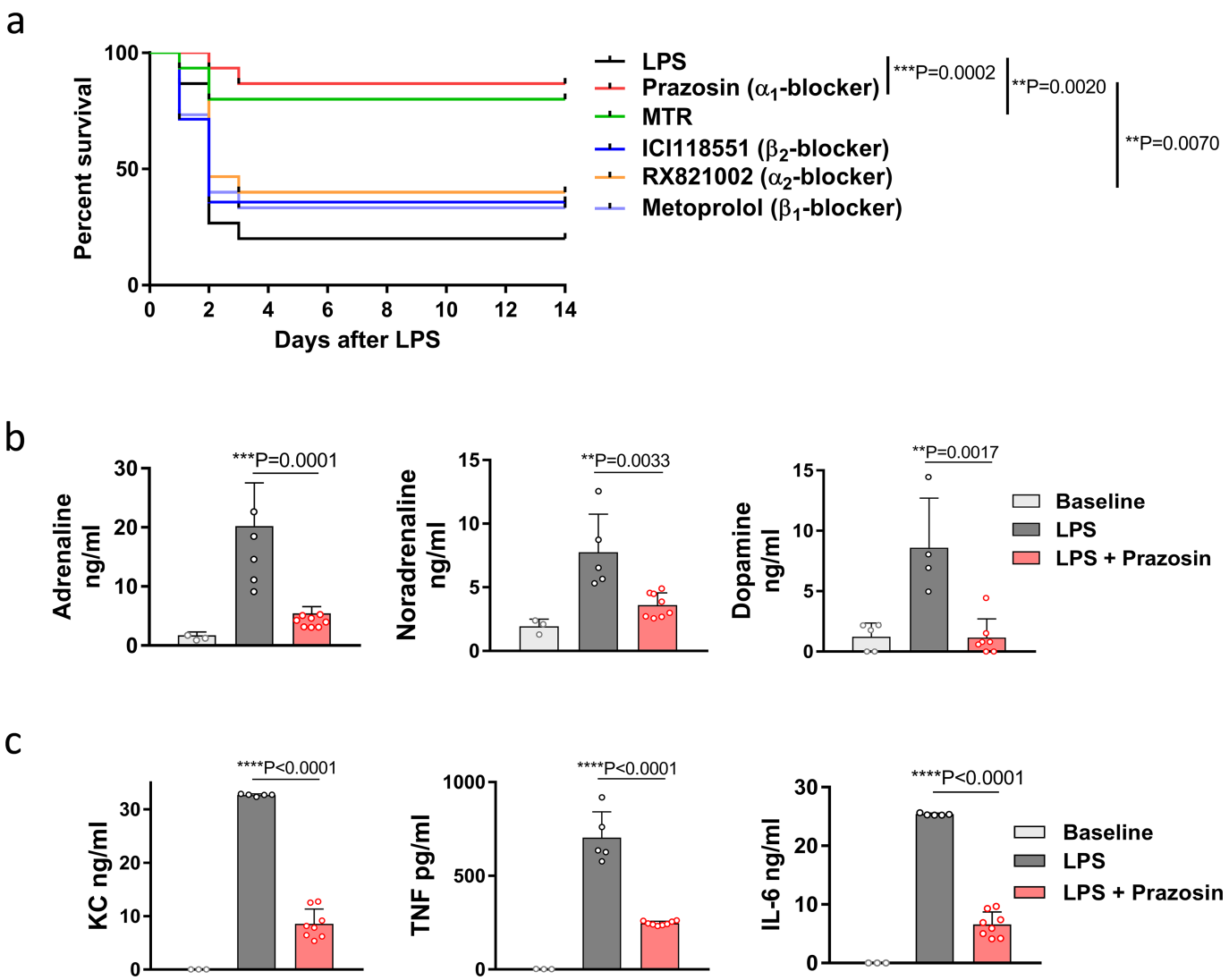
Extended Data Fig. 4 | Catecholamines modulate the cytokine release in macrophages in vitro. a, **b**, Human U937 macrophage-like cells were pre-treated with ANP or MTR for 10 min, then stimulated with LPS at 1 μg ml^{-1} and/or adrenaline at 15 ng ml^{-1}. Culture supernatants were analysed for catecholamines ($n = 3$ per column) (**a**) as well as the indicated cytokines ($n = 3$ per column) (**b**). **c**, *TH* expression of baseline and LPS-stimulated $Th^{+/+}$ or $Th^{\Delta LysM}$ macrophages ($n = 3$ per group), analysed by RT–PCR; results are normalized to ubiquitin C (*UBC*) expression. **d**, **e**, Supernatants of collected peritoneal macrophages from $Th^{+/+}$ or $Th^{\Delta LysM}$ mice, stimulated with LPS at 50 μg ml^{-1}, adrenaline 15 μg ml^{-1} or both for 24 h, were analysed for levels of adrenaline ($n = 3$), noradrenaline ($n = 3$) (**d**) and cytokines IL-6 ($n = 3$), KC ($n = 3$), MIP-2 ($n = 3$) and TNF ($n = 3$) (**e**). All data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test.

© 2018 Springer Nature Limited. All rights reserved.



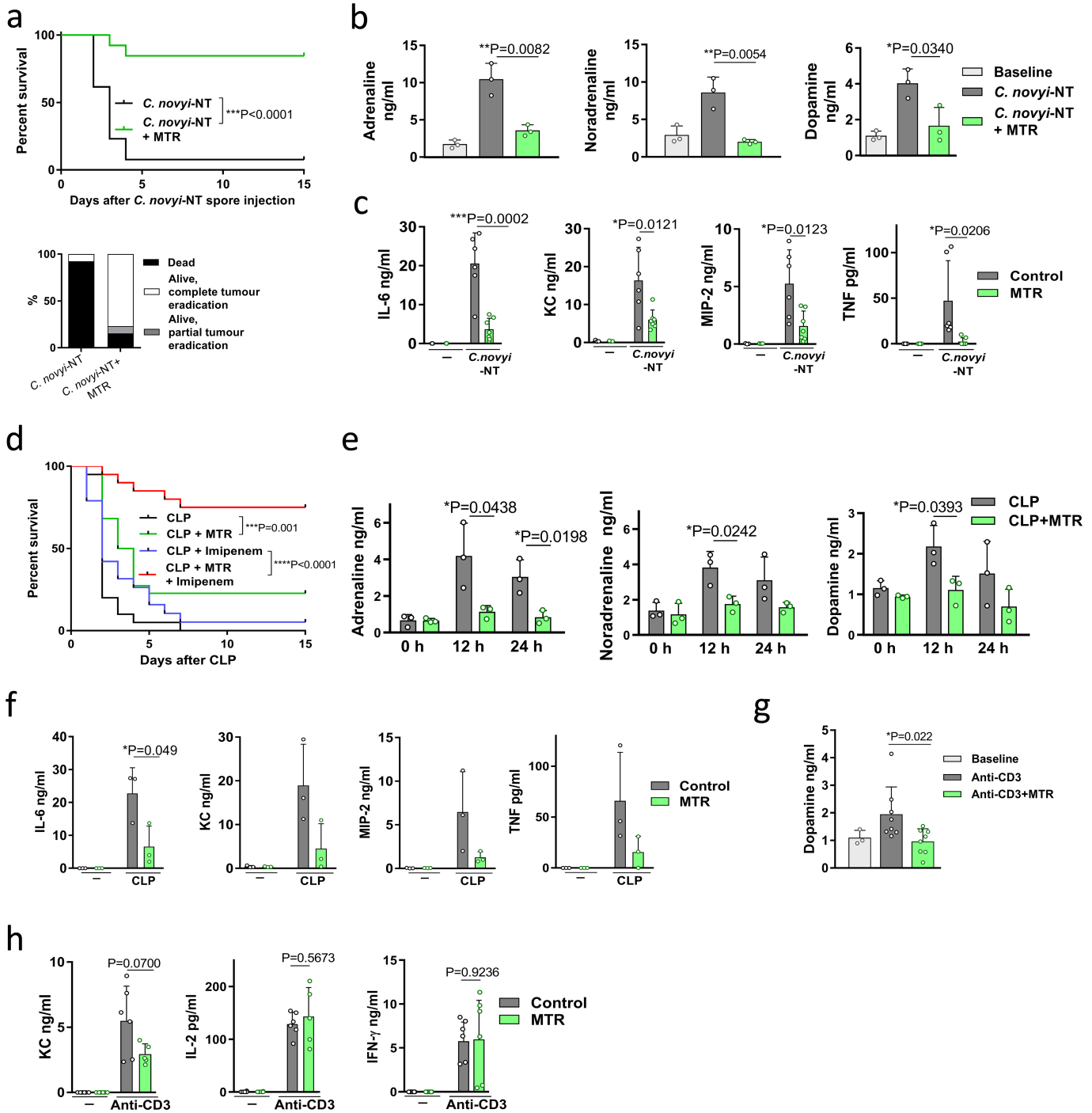
Extended Data Fig. 5 | Modulation of catecholamine synthesis by MTR dose-dependently determines survival and cytokine release. a, Survival of BALB/c mice stimulated with a lethal dose of LPS and treated with the indicated dose of MTR: MTR 20 mg kg^{-1} ($n = 5$ mice per group); MTR 30 mg kg^{-1} ($n = 10$ mice), MTR 40 mg kg^{-1} ($n = 12$) compared to LPS ($n = 10$ mice). Survival differences were analysed by two-sided log-rank test. **b**, **c**, Levels of plasma catecholamines ($n = 4$ per column) (**b**) and IL-6 ($n = 4$ per column), KC (left to right, $n = 4, 4, 3$ per column), IFN-γ ($n = 4$) and TNF ($n = 4, 4, 3$) (**c**) at different MTR doses measured 24 h after LPS injection. **d**, **e**, 24-h-time course of circulating adrenaline ($n = 5, 5, 5, 4, 5, 4, 5, 5$), noradrenaline ($n = 5$) and dopamine ($n = 5$) (**d**) and corresponding levels of IL-6 ($n = 4$), KC ($n = 7, 7, 7, 6, 5, 4, 5, 5$), IFN-γ ($n = 6, 6, 6, 8, 4, 8, 6, 4$) and TNF ($n = 6, 6, 6, 6, 6, 4, 7, 7$) (**e**) in LPS-treated mice receiving MTR 40 mg kg^{-1}. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**b**–**e**).

© 2018 Springer Nature Limited. All rights reserved.



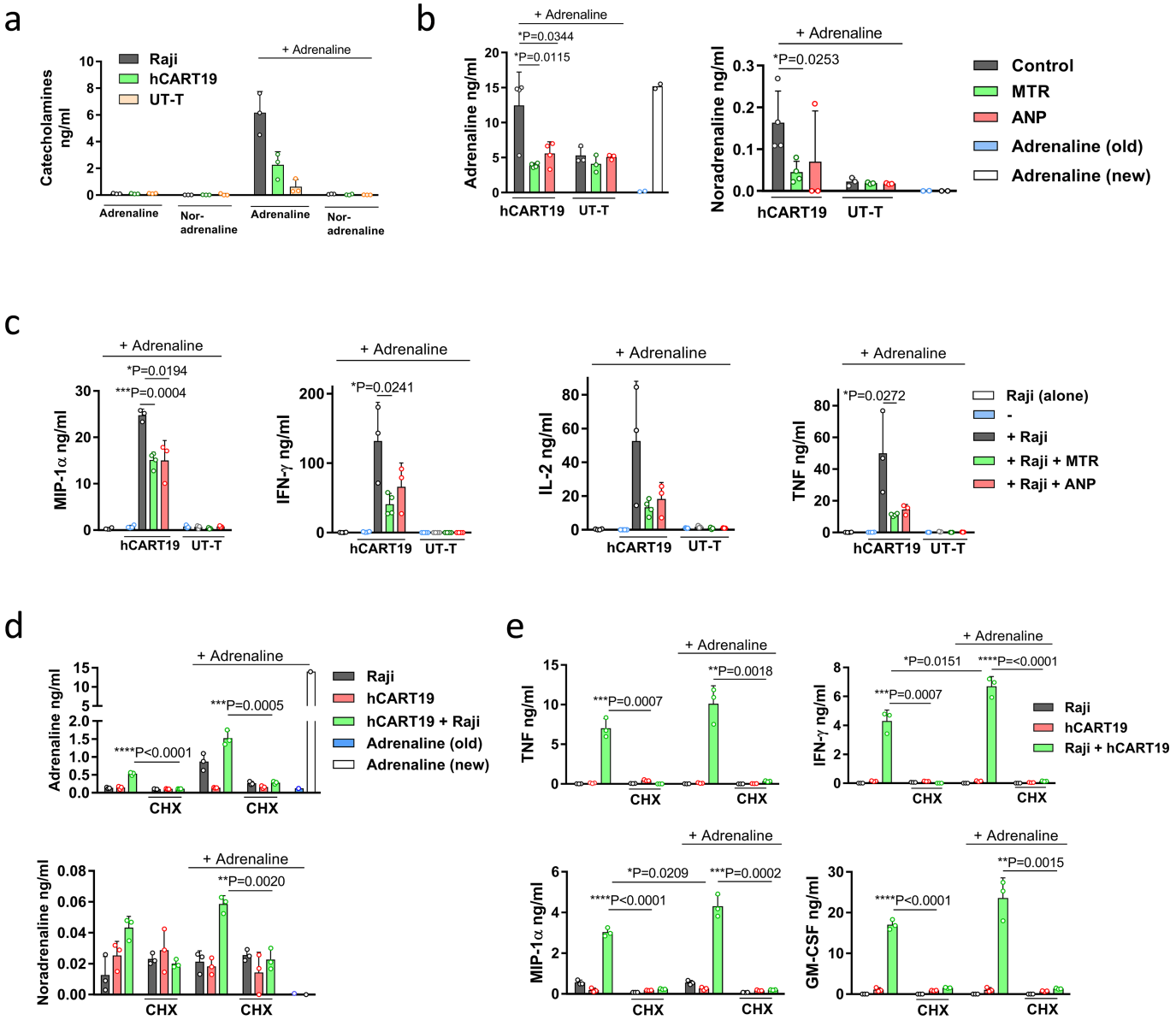
Extended Data Fig. 6 | Blockage of α_1-adrenoceptor mediates the survival in experimental systemic inflammatory syndrome. a, Kaplan–Meier curve of LPS-injected BALB/c mice treated with the indicated adrenoreceptor blockers ($n = 15$ animals per group). Survival differences were analysed by two-sided log-rank test. **b**, **c**, Levels of adrenaline, noradrenaline (left to right, $n = 3, 5, 8$ per column) and dopamine ($n = 5, 4, 7$) (**b**) as well as indicated cytokines ($n = 3, 5, 8$) (**c**) measured 24 h after LPS administration. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**b**, **c**).

© 2018 Springer Nature Limited. All rights reserved.



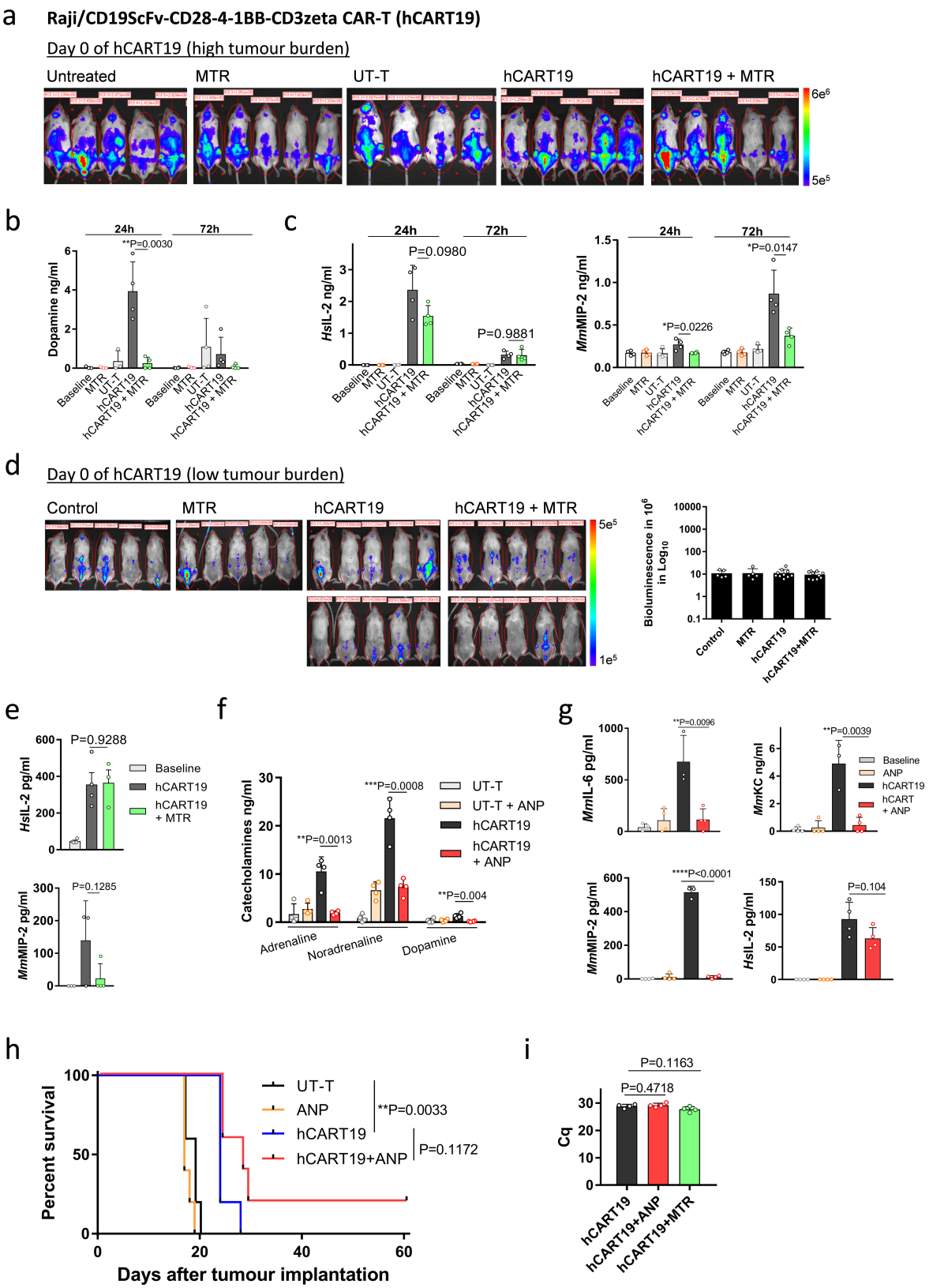
Extended Data Fig. 7 | Suppression of catecholamines with MTR reduces toxicity of oncolytic bacterium *C. novyi*-NT and polymicrobial sepsis. a, Survival (top panel) and therapeutic response (bottom panel) of CT26 tumour-bearing BALB/c mice undergoing *C. novyi*-NT treatment with or without MTR pre-treatment ($n = 13$ mice per group). Survival differences were analysed with two-sided log-rank test. **b**, **c**, Corresponding plasma levels of adrenaline ($n = 3$ independent samples per column), noradrenaline ($n = 3$), dopamine ($n = 3$) (**b**) and indicated cytokines (left to right, $n = 3, 3, 6, 7$ independent samples per column) (**c**), measured at baseline and 36 h after treatment. **d**, Survival of C57Bl/6 mice undergoing CLP, with the indicated treatments (CLP, $n = 20$ mice; MTR, $n = 22$; imipenem, $n = 19$; MTR + imipenem, $n = 20$ mice per group). Survival differences were analysed with two-sided log-rank test. **e**, Plasma levels of adrenaline ($n = 3$), noradrenaline ($n = 3$) and dopamine ($n = 3$) at the indicated time points after CLP, with or without MTR pre-treatment. **f**, Levels of indicated cytokines ($n = 3$) at baseline and 24 h after CLP, with or without MTR pre-treatment. **g**, **h**, Levels of plasma dopamine (left to right, $n = 3, 8, 8$ independent samples per column) (**g**) and KC ($n = 6, 6, 6, 5$), IL-2 ($n = 6, 6, 6, 5$) and IFN-γ ($n = 6$) (**h**) measured 24 h after α-CD3 treatment, with or without MTR. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**b**, **c**, **e**–**h**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Adrenaline stimulates cytokine release during hCART19–Raji cell interaction in vitro. a, Levels of catecholamines measured individually in supernatants of Raji cells ($n = 3$), hCART19 ($n = 3$) and UT-T ($n = 3$ per column) at baseline and when exposed to adrenaline. **b**, **c**, Co-cultures of hCART19 and Raji with or without MTR or ANP pre-treatment were stimulated with 15 ng ml^{-1} of adrenaline. Culture supernatants were collected after 24 h and analysed for adrenaline (left to right, $n = 4, 4, 4, 3, 3, 3, 2, 2$ per column) and noradrenaline ($n = 4, 4, 3, 3, 3, 3, 2, 2$). Adrenaline (old): adrenaline at 15 ng ml^{-1} was incubated at 37 °C for 24 h in the cell-free medium. Adrenaline (new): adrenaline at 15 ng ml^{-1} was added into the cell-free medium and immediately measured (**b**). Corresponding cytokine levels of MIP-1α ($n = 4, 4, 3, 4, 3, 3, 3, 3, 3$), IFN-γ ($n = 4, 4, 3, 4, 3, 4, 4, 4, 4$), IL-2 ($n = 4, 4, 3, 4, 3, 3, 3, 3, 3$) and TNF ($n = 4, 4, 3, 4, 3, 3, 3, 3, 3$) (**c**). UT-T served as control. **d**, **e**, As above, co-cultures of hCART19 and Raji with or without CHX were stimulated with 15 ng ml^{-1} of adrenaline in vitro. Levels of catecholamines ($n = 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1$) (**d**) and indicated human cytokines ($n = 3$) (**e**) were measured after 24 h. Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test.

© 2018 Springer Nature Limited. All rights reserved.

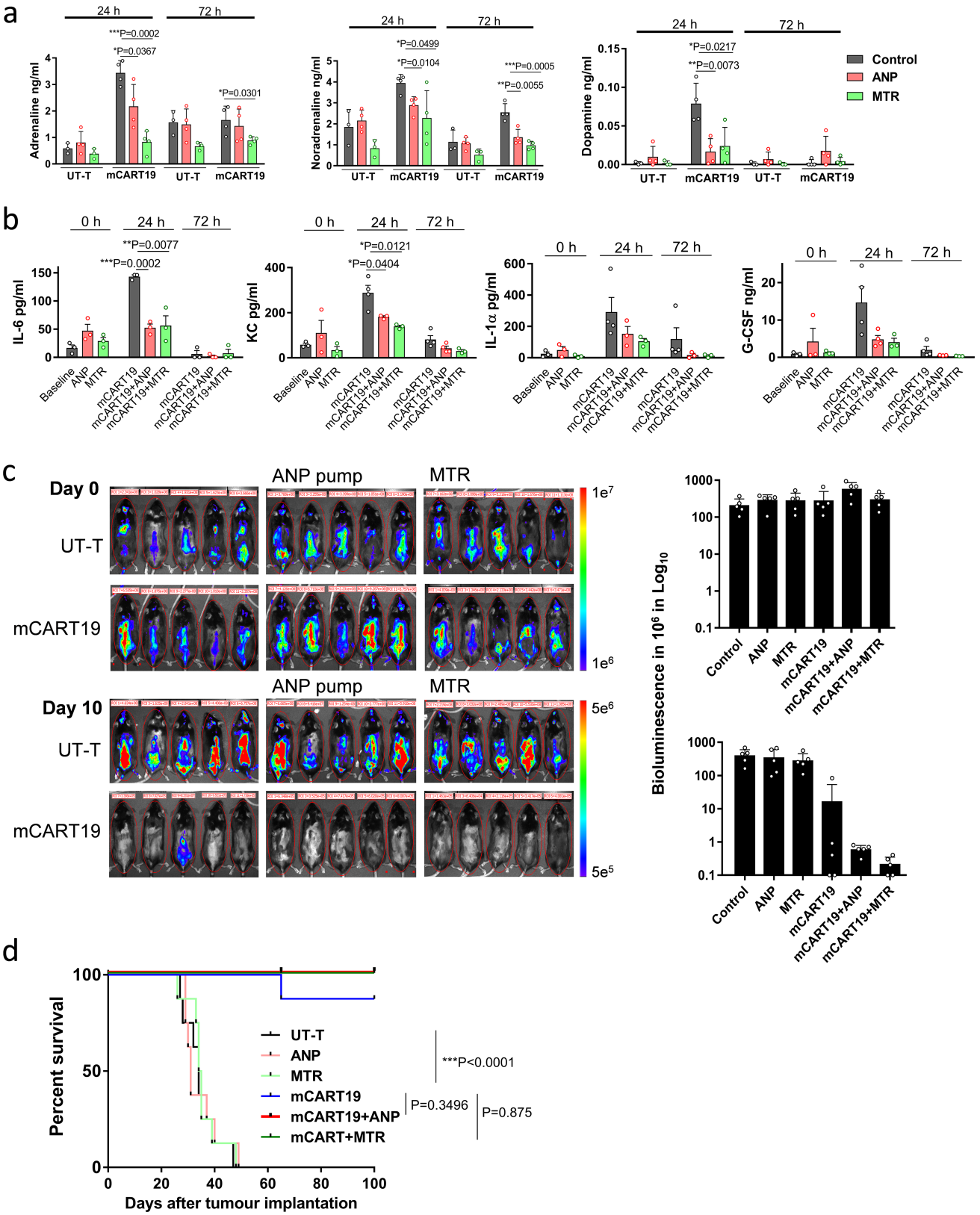


Extended Data Fig. 9 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | MTR and ANP prevent cytokine release in Raji/hCART19 mouse model. a, Bioluminescent images (BLI) of Raji-bearing NSGS mice with high tumour burden. At day 0, tumour engraftment was quantified by BLI and mice were assigned to the treatment groups (untreated, MTR, hCART19, hCART19+MTR, $n = 5$ mice per group; UT-T, $n = 4$). **b**, **c**, Levels of dopamine (left to right, $n = 3, 3, 3, 4, 4, 4, 3, 4, 4, 4$ per column) (**b**) and indicated cytokines ($n = 4$) (**c**) measured in mice (with high tumour burden) 24 and 72 h after hCART19 and UT-T injection. **d**, BLI of Raji-bearing NSGS mice with low tumour burden. At day 0, mice were randomly assigned based on tumour burden to receive hCART19, with or without MTR ($n = 10$ mice per group) or UT-T, with or without MTR ($n = 5$ mice per group). **e**, Levels of *Hs*IL-2 ($n = 4, 4, 3$) and *Mm*MIP-2 ($n = 3, 3, 4$) assessed 72 h after hCART19 injection in mice with low tumour burden. **f**, **g**, NSGS mice were injected with hCART19 4 days after Raji implantation and treated with ANP delivered via subcutaneously implanted osmotic pumps. Levels of circulating catecholamines ($n = 4$ per column) (**f**) and *Mm*IL-6, *Mm*KC and *Mm*MIP-2 ($n = 4, 4, 3, 4$) as well as *Hs*IL-2 ($n = 4$) (**g**) were assessed 24 h after hCART19 administration. **h**, Survival of Raji cell-bearing NSGS mice treated with hCART19 and ANP ($n = 5$ per group); analysed by two-sided log-rank test. **i**, Level of circulating hCART19 10 days after treatment, determined by C_q from RT–PCR and analysed in triplicates ($n = 4$ per group). Data are presented as mean ± s.d. with individual data points shown, analysed by two-tailed *t*-test (**b**, **c**, **e**–**i**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | MTR and ANP prevent cytokine release in syngeneic Eμ-ALL model without compromising anti-tumour efficacy. **a**, **b**, Circulating catecholamines (left to right, $n = 3, 4, 3, 4, 4, 4, 3, 4, 3, 4, 4, 4$ per column/graph) (**a**) and murine cytokines IL-6 ($n = 3$ per column), KC ($n = 3, 3, 3, 4, 3, 3, 4, 4, 3$ per column), IL-1α ($n = 3, 3, 3, 4, 3, 3, 4, 3, 3$ per column) and G-CSF ($n = 3, 3, 3, 4, 4, 3, 4, 3, 3$ per column) (**b**), assessed at 24 and 72 h after mCART19 injection. Data are presented as means ± s.d. with individual data points shown, analysed by two-tailed *t*-test. **c**, BLI was performed before and 10 days after mCART19 cell injection, with or without ANP and MTR pre-treatment ($n = 5$ animals per group). BLI radiance was used to quantify the tumour burden during the treatment course (right). **d**, Percentage survival of Eμ-ALL-mice after mCART19 cell transfer ($n = 8$ mice per group). Survival differences were analysed by two-sided log-rank test.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0750-6

NP220 mediates silencing of unintegrated retroviral DNA

Yiping Zhu[1,2,3,4], Gary Z. Wang[2,3,4], Oya Cingöz[2,3] & Stephen P. Goff[1,2,3]*

The entry of foreign DNA into many mammalian cell types triggers the innate immune system, a complex set of responses to prevent infection by pathogens. One aspect of the response is the potent epigenetic silencing of incoming viral DNAs[1], including the extrachromosomal DNAs that are formed immediately after infection by retroviruses. These unintegrated viral DNAs are very poorly transcribed in all cells, even in permissive cells, in contrast to the robust expression that is observed after viral integration[2–5]. The factors that are responsible for this low expression have not yet been identified. Here we performed a genome-wide CRISPR–Cas9 screen for genes that are required for silencing an integrase-deficient MLV–GFP reporter virus to explore the mechanisms responsible for repression of unintegrated viral DNAs in human cells. Our screen identified the DNA-binding protein NP220, the three proteins (MPP8, TASOR and PPHLN1) that comprise the HUSH complex—which silences proviruses in heterochromatin[6] and retrotransposons[7,8]—the histone methyltransferase SETDB1, and other host factors that are required for silencing. Further tests by chromatin immunoprecipitation showed that NP220 is the key protein that recruits the HUSH complex, SETDB1 and the histone deacetylases HDAC1 and HDAC4 to silence the unintegrated retroviral DNA. Knockout of *NP220* accelerates the replication of retroviruses. These experiments identify the molecular machinery that silences extrachromosomal retroviral DNA.

Retroviral infections begin with the reverse transcription of the RNA genome in the cytoplasm to form a linear double-stranded DNA that is soon delivered into the nucleus[9]. A portion of this linear DNA gives rise to two circular DNA forms that contain one or two tandem copies of the long terminal repeat (LTR) sequences at the ends of the linear DNA[10,11]. The linear DNA is inserted into the host genome to form the integrated provirus, which is actively transcribed and produces progeny virus. By contrast, the unintegrated DNAs are transcriptionally silent, do not replicate and disappear over time. To analyse the mechanism of silencing of unintegrated retroviral DNAs, we infected HeLa cells with integrase-deficient (IN(D184A)) or integrase-proficient (IN(WT)) MLV-Luc viruses and monitored expression of the luciferase reporter. Comparable levels of retroviral DNA were produced, but expression of the reporter was approximately 30-fold lower in cells infected with the IN(D184A) virus than those infected with the IN(WT) virus (Extended Data Fig. 1a, b). Treatment of the cells with the histone deacetylase (HDAC) inhibitor trichostatin A (TSA) markedly increased luciferase expression of unintegrated IN(D184A) MLV-Luc DNA without measureable effect on the integrated wild-type control virus (Extended Data Fig. 1c, d). Chromatin immunoprecipitation (ChIP) experiments showed that H3 histones on unintegrated viral DNA were largely deacetylated and carried repressive H3K9me3 marks but not H3K27me3 marks. By contrast, H3 histones on integrated retroviral DNA were heavily acetylated and barely methylated at H3K9 or H3K27 (Extended Data Fig. 1e–j).

To identify host factors responsible for the silencing of unintegrated retroviral DNA, we performed an unbiased genome-wide CRISPR–Cas9 screen, selecting for the knockout of host factor genes that relieve the silencing (Fig. 1a). HeLa-Cas9 cells transduced with a CRISPR single-guide RNA (sgRNA) library were infected with an integrase-deficient MLV-GFP reporter virus and GFP$^+$ cells were isolated by sorting (Fig. 1b). After a second round of selection, targeted genes, the knockout of which was increased in the selected cells, were identified by sequencing of sgRNAs. Five genes stood out: *NP220* (also known as *ZNF638*), all three subunits of the HUSH complex (*MPP8* (also known as *MPHOSPH8*), *TASOR* (also known as *FAM208A*) and

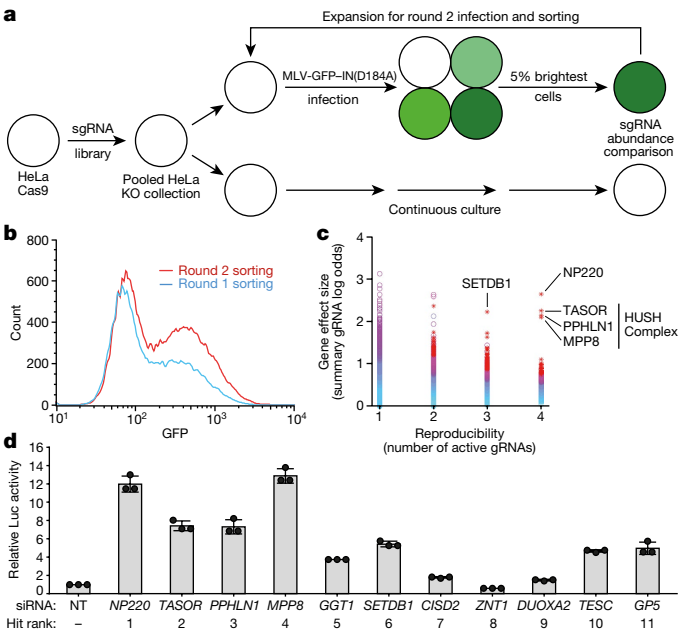


Fig. 1 | CRISPR–Cas9 screen to identify host factors responsible for silencing of unintegrated retroviral DNA. a, Flowchart of the genome-wide CRISPR–Cas9 screening strategy. A pooled collection of HeLa knockout (KO) cells was infected with an integrase-deficient MLV-GFP virus and the 5% brightest GFP$^+$ cells were isolated by sorting. These cells were subjected to a second round of infection and selection and DNAs were recovered for analysis of the resident sequences that encode the CRISPR single-guide RNAs (sgRNAs). **b**, GFP signals detected by fluorescence-activated cell sorting during round 1 and round 2 sorting. **c**, Candidate genes that are essential for the silencing were identified by analysing the fold change in abundance over control and the number of enriched sgRNAs per gene using HiTSelect software. **a**–**c**, The CRISPR screen (**a**), GFP monitoring experiment (**b**) and sgRNA analysis (**c**) were performed only once. **d**, Validation of candidate genes from the screen. HeLa cells were transfected with the indicated siRNAs and then infected with an integrase-deficient (IN(D184A)) MLV-Luc virus. Luciferase activities were measured and activity in cells transfected with non-targeting (NT) siRNA was set to 1. Data are mean ± s.d.; $n = 3$ independent experiments.

[1]Howard Hughes Medical Institute, Columbia University, New York, NY, USA. [2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [3]Department of Microbiology and Immunology, Columbia University, New York, NY, USA. [4]These authors contributed equally: Yiping Zhu, Gary Z. Wang. *e-mail: spg1@cumc.columbia.edu

© 2018 Springer Nature Limited. All rights reserved.

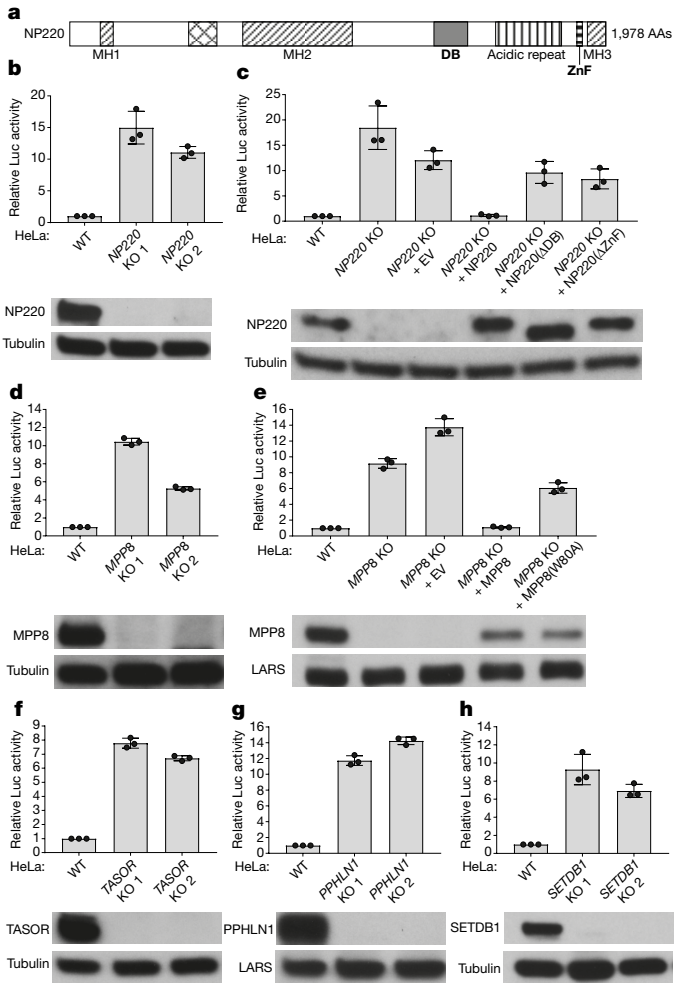


Fig. 2 | NP220, HUSH complex and SETDB1 are required for the silencing of unintegrated MLV DNA. a, Schematic domain organization of NP220. MH1, MH2, MH3, domains that are homologous to matrin 3; RS, arginine- and serine-rich domain; DB, DNA-binding domain; ZnF, Zinc-finger motif. **b**–**h**, NP220 (**b**, **c**), MPP8 (**d**, **e**), TASOR (**f**), PPHLN1 (**g**) and SETDB1(**h**) are required for the silencing of unintegrated MLV DNA. Indicated cells were infected with an integrase-deficient (IN(D184A)) MLV-Luc virus. Luciferase activities were measured and activity in parental (wild-type (WT)) HeLa cells was set to 1 (top). The expression of indicated proteins was determined by western blot (bottom). EV, empty vector; NP220(ΔDB), NP220 in which the DNA-binding domain is deleted; NP220(ΔZnF), NP220 in which the zinc-finger domain is deleted; MPP8(W80A), MPP8 in which a mutation (W80A) renders the protein deficient in H3K9me3 binding activity. Data are mean ± s.d.; $n = 3$ independent experiments.

PPHLN1)) as well as *SETDB1*, a histone methyl transferase (HMT) (Fig. 1c and Supplementary Table 1).

NP220 is a nuclear double-stranded DNA (dsDNA)-binding protein with a preference for cytidine clusters[12] and contains a DNA-binding domain (DB) and a single C-terminal C_2H_2-type zinc finger (ZnF) motif (Fig. 2a). Knockdown or knockout of NP220 resulted in a marked increase in expression of luciferase from unintegrated MLV DNA (Figs 1d, 2b and Extended Data Fig. 2a, b) without affecting viral DNA levels (Extended Data Fig. 1k). Re-expression of wild-type NP220, but not deletion mutants that lacked either the DNA-binding domain (NP220(ΔDB)) or the zinc finger motif (NP220(ΔZnF)), restored silencing in *NP220* knockout cells (Fig. 2c). Therefore, both the DNA-binding domain and the zinc finger of NP220 are required for silencing. The HUSH complex was previously identified to have a role in silencing proviruses that are integrated into heterochromatin[6] and also evolutionarily young retrotransposons[7,8]. Knockout of the HUSH complex subunits—*MPP8*, *TASOR* or *PPHLN1*—similarly increased reporter expression from unintegrated DNA (Fig. 2d–g). The MPP8 subunit is known to exhibit methyl H3K9-binding activity and tryptophan 80 (W80) is critical for this binding[6,13,14]. Re-expression of wild-type MPP8, but not a mutant with a W80A substitution (MPP8(W80A)), restored the repression of unintegrated DNA in the *MPP8* knockout line (Fig. 2e). SETDB1 is an HMT that is responsible for generating H3K9me3 marks. Knockout or knockdown of *SETDB1*, but not other HMTs (*SETDB2*, *SUV39H1*, *SUV39H2*, *EHMT1*, *EHMT2* or *EZH2*) again relieved the silencing of unintegrated retroviral DNA (Fig. 2h and Extended Data Fig. 2c, d).

Co-immunoprecipitation experiments were used to analyse the interaction of NP220 with the HUSH complex in HeLa cells. Immunoprecipitation of NP220 resulted in efficient co-immunoprecipitation of MPP8 and TASOR (Fig. 3a and Extended Data Fig. 3a), independent of DNA or RNA (Extended Data Fig. 4b). The N-terminal 471 amino acids of NP220 were sufficient to mediate the co-immunoprecipitation with MPP8 (Extended Data Fig. 3b). In *MPP8* knockout cells, NP220 did not co-immunoprecipitate either MPP8 or TASOR (Fig. 3a), consistent with MPP8 serving as the direct partner of NP220. In *TASOR* knockout cells, levels of both TASOR and MPP8 were very low (Fig. 3a) and neither was detected to be interacting with NP220 (Fig. 3a). In *PPHLN1* knockout cells, the levels of MPP8 and TASOR were increased (Fig. 3a) and immunoprecipitation of NP220 resulted in correspondingly increased co-immunoprecipitation of both MPP8 and TASOR (Fig. 3a).

ChIP assays were performed to monitor binding of these proteins to unintegrated viral DNA. NP220 was able to bind to both linear and circular unintegrated DNA (Fig. 3b and Extended Data Fig. 3e). *NP220* knockout eliminated the ChIP signal (Fig. 3b) and re-expression of wild-type NP220 or the NP220(ΔZnF) mutant restored binding, whereas expression of the NP220(ΔDB) mutant did not (Extended Data Fig. 3c). Notably, NP220 bound to DNA even when expression of any of the HUSH subunits or SETDB1 was eliminated, suggesting that NP220 is the primary DNA-binding component (Fig. 3b). We also detected robust binding of each subunit of the HUSH complex (MPP8, TASOR and PPHLN1) to unintegrated viral DNA in wild-type cells, but not in *MPP8* or *SETDB1* knockout cells (Fig. 3c–e). Binding of the HUSH complex to unintegrated viral DNA required histone methylation, because re-expression of MPP8(W80A), which lacks methyl-binding activity, in the *MPP8* knockout line showed only weak binding to viral DNA (Extended Data Fig. 3d). None of the HUSH subunits or SETDB1 bound to viral DNA in *NP220* knockout cells (Fig. 3c–f), suggesting that NP220 has a key role in bringing the HUSH complex and SETDB1 to DNA.

Knockout of *NP220*, *MPP8* or *SETDB1* significantly decreased H3K9me3 modification on unintegrated retroviral DNA (Fig. 3g), indicating that NP220, the HUSH complex and SETDB1 are all required for H3K9 trimethylation on unintegrated retroviral DNA. Knockdown of *HDAC1* or *HDAC4* also increased the expression of unintegrated retroviral DNA (Fig. 4a, b) and increased the levels of acetylated histone H3 (Fig. 4c). Co-immunoprecipitation assays showed that endogenous NP220 bound to HDAC4, but not other HDACs (Fig. 4d and Extended Data Fig. 4a). The NP220–HDAC4 interaction was independent of DNA or RNA (Extended Data Fig. 4b). NP220(ΔZnF) did not co-immunoprecipitate with HDAC4 (Fig. 4e), indicating that the C-terminal zinc finger motif is critical for this interaction. ChIP assays showed that both HDAC1 and HDAC4 were bound to unintegrated retroviral DNA and that binding was reduced in *NP220* knockout cells (Fig. 4f, g). Re-expression of wild-type NP220 in *NP220* knockout cells decreased the levels of acetylated H3 on unintegrated DNA, whereas NP220(ΔDB) and NP220(ΔZnF) failed to do so (Fig. 4h). These results indicate that NP220 utilizes its C-terminal zinc finger to recruit HDAC4, and probably HDAC1, to deacetylate histone H3 and thereby silence the expression of unintegrated retroviral DNA. Deacetylation appears to be mechanistically upstream of H3 methylation: HDAC inhibition and H3 acetylation led to decreased H3K9me3 (Extended

© 2018 Springer Nature Limited. All rights reserved.

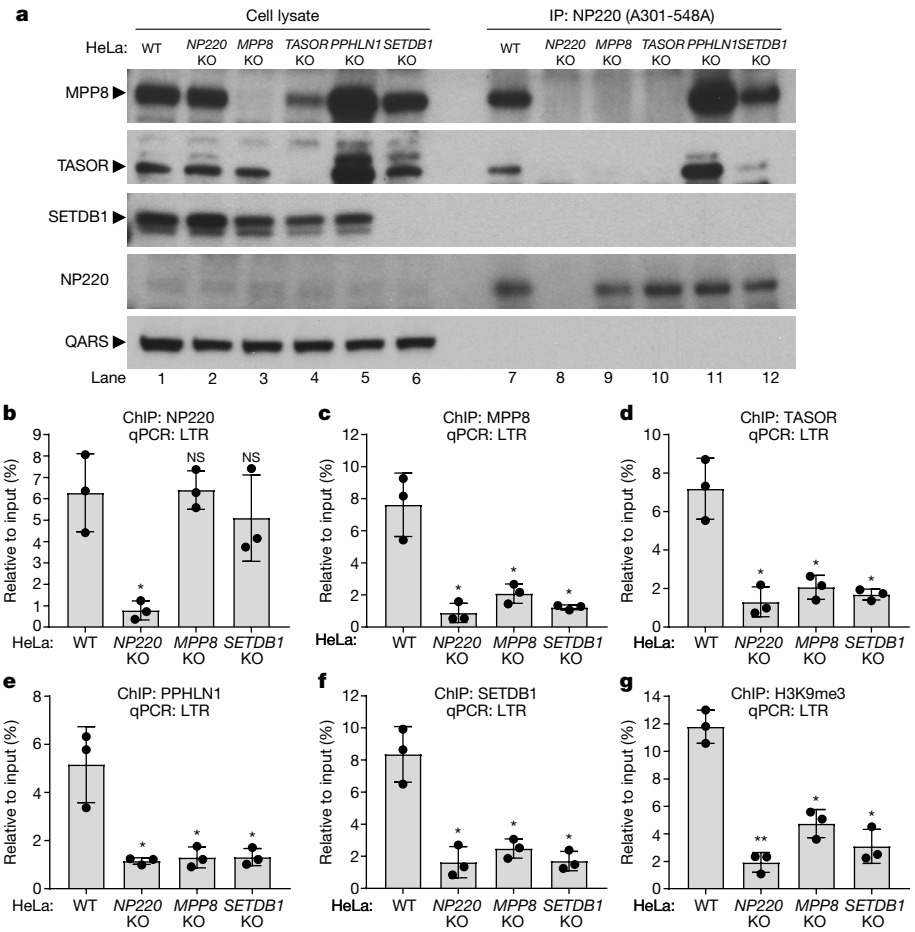


Fig. 3 | NP220 recruits the HUSH complex and SETDB1 to silence unintegrated MLV DNA. a, NP220 interacts with the HUSH complex. Endogenous NP220 was immunoprecipitated (IP) from lysates of the indicated HeLa cell lines and co-immunoprecipitating proteins were analysed by western blot. Images are representative of two independent experiments with similar results. **b**–**g**, NP220 recruits the HUSH complex and SETDB1 to deposit H3K9me3 on unintegrated MLV DNA. Indicated HeLa cell lines were infected with an integrase-deficient (IN(D184A)) MLV-Luc virus. ChIP was performed using antibodies against NP220 (**b**), MPP8 (**c**), TASOR (**d**), PPHLN1 (**e**), SETDB1 (**f**), H3K9me3 (**g**), followed by qPCR using primers that target the LTR. qPCR data from each ChIP experiment were calculated as the percentage of input DNA. Data are mean ± s.d.; $n = 3$ independent experiments. NS, not significant ($P > 0.05$); $^{*}P < 0.05$; $^{**}P < 0.01$. P values are from paired two-sided Student's t-tests. Exact P values are included in the Source Data associated with this figure.

Data Fig. 4c), whereas *SETDB1* knockout and H3K9 demethylation did not lead to H3 acetylation (Extended Data Fig. 4d).

We next examined silencing of retroviruses other than MLV. Knockout or knockdown of *NP220* increased the expression of unintegrated DNA of human immunodeficiency virus type 1 (HIV-1) and Mason–Pfizer monkey virus (MPMV), but not Rous sarcoma virus (RSV) (Extended Data Fig. 5a–e). Knockout of the components of the HUSH complex had no or minimal effect on the silencing of unintegrated HIV-1, MPMV or RSV viral DNA (Extended Data Fig. 5c–e). Knockdown of *HDAC1* and/or *HDAC4* relieved the silencing of unintegrated HIV-1 and MPMV DNA (Extended Data Fig. 5f, g). We conclude that the silencing of unintegrated DNAs of the various retrovirus genera is mediated by a distinctive variety of host factors.

NP220 preferentially binds to cytidine clusters in dsDNA at the consensus sequence of CCCCC(G/C)[12]. The MLV U3 promoter region contains five close matches to consensus NP220-binding sites (Fig. 5a and Extended Data Fig. 6a). By electrophoretic mobility shift assays, incubation of the NP220 DNA-binding domain (NP220 DB) with a 84-bp biotin-labelled DNA fragment of the MLV LTR that contained two potential NP220-binding sites (biotin–DNA84) produced a mobility shifted band, and the resulting shift was sensitive to competition by cold probe (DNA84), but not by a mutant probe in which the NP220 binding sites were deleted (DNA72) (Fig. 5b). These results indicate that NP220 specifically binds to the cytidine clusters in MLV DNA. To test the importance of these DNA sequences in NP220-mediated silencing of unintegrated DNA, we generated a panel of variant reporters (Fig. 5a). Mutations M1 and M2 that delete the first three NP220-binding sites had no measureable effect on the LTR promoter activity when integration was allowed, but allowed higher expression of unintegrated DNAs (Extended Data Fig. 6b, c). Expression of unintegrated MLV DNA in which the U3 region was replaced by RSV U3—which lacks cytidine clusters—or in which the NP220-binding sites were deleted or mutated, was less affected by *NP220* knockout (Fig. 5c and Extended Data Fig. 6c). ChIP assays showed that the association of NP220 with viral DNA—similar to the silencing of expression—decreased upon the replacement, deletion or mutation of the NP220-binding DNA sequences (Fig. 5d). In the case of HIV-1, we identified five consensus NP220-binding sequences in the U3 promoter region (Extended Data Fig. 7a, b). Deletion of these putative binding sequences made unintegrated HIV-1 DNA less responsive to *NP220* knockout (Extended Data Fig. 7c, d). It should be noted that the last three binding sequences overlap with SP1-binding sites and deletion of these sequences severely diminished HIV-1 LTR basal promoter activity (Extended Data Fig. 7e).

NP220 not only silenced non-integrating viral DNA vectors, but also influenced infection by integration-competent vectors and even replicating viruses. MLV DNA was silenced, marked with histone deacetylation and bound by NP220, and *NP220* knockout markedly increased the expression of viral DNA at 12 h after infection, when most of the viral DNA is unintegrated, but not 48 h after infection, when most of the viral

© 2018 Springer Nature Limited. All rights reserved.

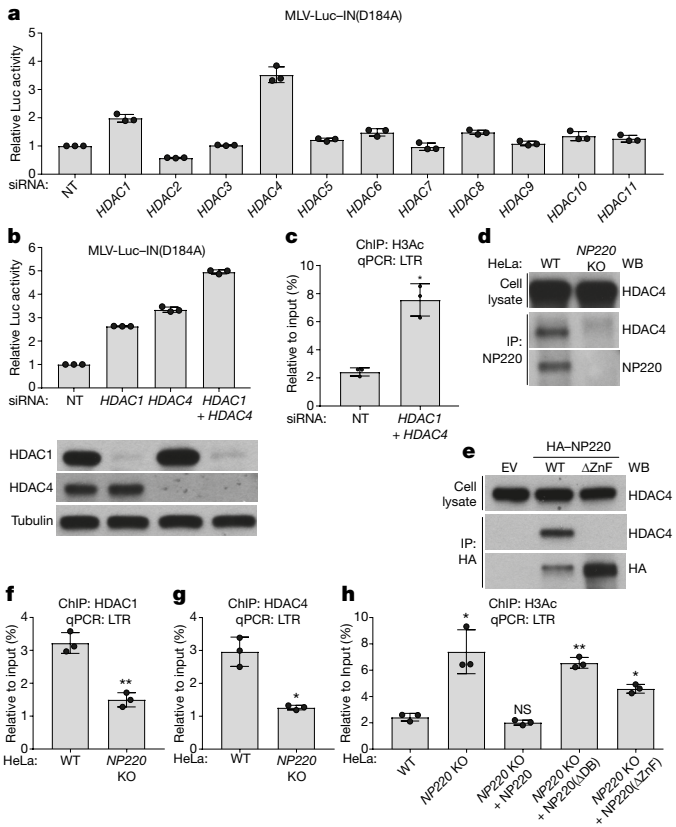


Fig. 4 | NP220 recruits HDAC1 and HDAC4 to deacetylate histone H3 on unintegrated retroviral DNA. a–c, HDAC1 and HDAC4 are required for the silencing of unintegrated retroviral DNA. HeLa cells were transfected with indicated siRNAs and then infected with an integrase-deficient (IN(D184A)) MLV-Luc virus. **a**, **b**, Luciferase activities were measured and luciferase activity of non-targeting (NT) siRNA-transfected cells was set to 1. The expression of HDAC1 and HDAC4 was determined by western blot (**b**, bottom). **c**, ChIP was performed using antibodies against pan-acetyl H3 followed by qPCR using primers that target the LTR. **d**, **e**, NP220 interacts with HDAC4. **d**, Endogenous NP220 was immunoprecipitated from lysates of the indicated HeLa cell lines. **e**, Haemagglutinin (HA)-tagged NP220 or NP220 in which the zinc finger domain was deleted (ΔZnF) were introduced into *NP220* knockout HeLa cells. NP220 was then immunoprecipitated from HeLa cell lysates using an anti-haemagglutinin antibody. The co-immunoprecipitating protein HDAC4 was analysed by western blot. Images are representative of two independent experiments with similar results. **f–h**, NP220 recruits HDAC1 and HDAC4 to deacetylate histone H3 on unintegrated retroviral DNA. Indicated cells were infected with an integrase-deficient (IN(D184A)) MLV-Luc virus. ChIP was performed using antibodies against HDAC1 (**f**), HDAC4 (**g**), pan-acetyl H3 (**h**), followed by qPCR using primers that target the LTR. **a–c**, **f–h**, Data are mean ± s.d.; $n = 3$ independent experiments. NS, not significant ($P > 0.05$); *$P < 0.05$; **$P < 0.01$. *P* values are from paired two-sided Student's *t*-tests. Exact *P* values are included in the Source Data associated with this figure.

DNA is integrated into the host genome (Fig. 5e, f and Extended Data Fig. 6d–g). The rate at which MLV and HIV-1 spread in *NP220* KO cells was faster than in control cells (Fig. 5g, h and Extended Data Fig. 7f).

Here we define the mechanisms and the machinery by which silencing is imposed on unintegrated retroviral DNAs (Extended Data Fig. 8). These findings define new functions for NP220 and the HUSH complex. The restriction is sufficiently strong that many viruses have evolved means to evade or inactivate this machinery: the Vpr and Vpx proteins from primate immunodeficiency viruses mediate degradation of the HUSH complex to relieve silencing of proviruses[15,16], and the ICP0 protein of herpes simplex virus type 1 relieves HDAC-mediated silencing of viral DNA[17,18]. The silencing mechanisms and machinery

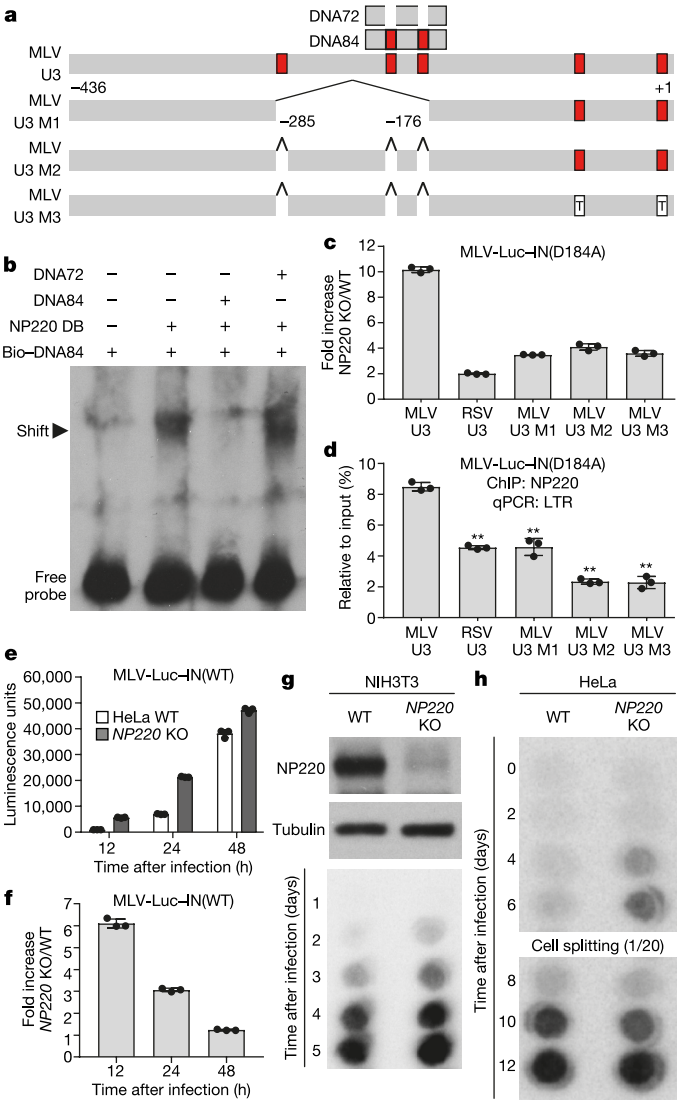


Fig. 5 | The binding specificity of NP220 to unintegrated retroviral DNA. a, Locations of putative NP220 binding sites in the MLV U3 region. Putative NP220-binding sites are indicated in red. T, C nucleotides mutated to T. **b**, NP220 binds a DNA fragment from the MLV U3 region. Indicated DNA fragments were incubated with the NP220 DNA-binding domain (NP220 DB) and shifted bands were probed with an antibody that recognized biotin. **c**, **d**, Specific DNA sequences are responsible for NP220 silencing and binding of unintegrated retroviral DNA. Indicated cells were infected with integrase-deficient (IN(D184A)) MLV-Luc viruses bearing the indicated deletions or mutations in the U3 region. **c**, Luciferase activities were measured and the fold increase (*NP220* knockout/*NP220* wild type) was calculated as the ratio of luciferase activity in knockout cells compared to wild-type cells. **d**, ChIP was performed to assess the association of NP220 with unintegrated MLV-Luc DNA. RSV U3, replacement of MLV U3 with RSV U3. **$P < 0.01$. *P* values are from paired two-sided Student's *t*-tests. Exact *P* values are included in the Source Data associated with this figure. **e**, **f**, Indicated HeLa cells were infected with an integrase-proficient (IN(WT)) MLV-Luc virus. **e**, At the indicated times after infection, luciferase activities were measured. **f**, Fold increase (*NP220* knockout/*NP220* wild type) was calculated as the ratio of luciferase activity in knockout cells compared to wild-type cells. **g**, **h**, Knockout of *NP220* increases the rate of MLV replication in a spreading infection. **g**, Top, the expression of NP220 was determined by western blot. **g**, **h**, Indicated cells were infected with replication-competent ecotropic (**g**) or amphotropic (**h**) MLV. Viral spreading was monitored using an assay for reverse transcriptase in the culture medium. **c–f**, Data are mean ± s.d.; $n = 3$ independent experiments. **b**, **g**, **h**, Images are representative of two independent experiments with similar results.

© 2018 Springer Nature Limited. All rights reserved.

uncovered here may have wide-ranging implications for the design of non-integrating retroviral vectors for gene therapy and induction of gene expression after plasmid DNA transfection.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0750-6.

Received: 8 May 2018; Accepted: 22 October 2018;
Published online 28 November 2018.

1. Orzalli, M. H. & Knipe, D. M. Cellular sensing of viral DNA and viral evasion mechanisms. *Annu. Rev. Microbiol.* **68**, 477–492 (2014).
2. Schwartzberg, P., Colicelli, J. & Goff, S. P. Construction and analysis of deletion mutations in the pol gene of Moloney murine leukemia virus: a new viral function required for productive infection. *Cell* **37**, 1043–1052 (1984).
3. Schneider, W. M., Wu, D. T., Amin, V., Aiyer, S. & Roth, M. J. MuLV IN mutants responsive to HDAC inhibitors enhance transcription from unintegrated retroviral DNA. *Virology* **426**, 188–196 (2012).
4. Sakai, H. et al. Integration is essential for efficient gene expression of human immunodeficiency virus type 1. *J. Virol.* **67**, 1169–1174 (1993).
5. Nakajima, N., Lu, R. & Engelman, A. Human immunodeficiency virus type 1 replication in the absence of integrase-mediated DNA recombination: definition of permissive and nonpermissive T-cell lines. *J. Virol.* **75**, 7944–7955 (2001).
6. Tchasovnikarova, I. A. et al. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. *Science* **348**, 1481–1485 (2015).
7. Liu, N. et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature* **553**, 228–232 (2018).
8. Robbez-Masson, L. et al. The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes. *Genome Res.* **28**, 836–845 (2018).
9. Varmus, H. Retroviruses. *Science* **240**, 1427–1435 (1988).
10. Yoshimura, F. K. & Weinberg, R. A. Restriction endonuclease cleavage of linear and closed circular murine leukemia viral DNAs: discovery of a smaller circular form. *Cell* **16**, 323–332 (1979).
11. Kilzer, J. M. et al. Roles of host cell factors in circularization of retroviral DNA. *Virology* **314**, 460–467 (2003).
12. Inagaki, H. et al. A large DNA-binding nuclear protein with RNA recognition motif and serine/arginine-rich domain. *J. Biol. Chem.* **271**, 12525–12531 (1996).
13. Chang, Y., Horton, J. R., Bedford, M. T., Zhang, X. & Cheng, X. Structural insights for MPP8 chromodomain interaction with histone H3 lysine 9: potential effect of phosphorylation on methyl-lysine binding. *J. Mol. Biol.* **408**, 807–814 (2011).
14. Kokura, K., Sun, L., Bedford, M. T. & Fang, J. Methyl-H3K9-binding protein MPP8 mediates E-cadherin gene silencing and promotes tumour cell motility and invasion. *EMBO J.* **29**, 3673–3687 (2010).
15. Chougui, G. et al. HIV-2/SIV viral protein X counteracts HUSH repressor complex. *Nat. Microbiol.* **3**, 891–897 (2018).
16. Yurkovetskiy, L. et al. Primate immunodeficiency virus proteins Vpx and Vpr counteract transcriptional repression of proviruses by the HUSH complex. *Nat. Microbiol.* https://doi.org/10.1038/s41564-018-0256-x (2018).
17. Gu, H. & Roizman, B. The two functions of herpes simplex virus 1 ICP0, inhibition of silencing by the CoREST/REST/HDAC complex and degradation of PML, are executed in tandem. *J. Virol.* **83**, 181–187 (2009).
18. Poon, A. P., Gu, H. & Roizman, B. ICP0 and the US3 protein kinase of herpes simplex virus 1 independently block histone deacetylation to enable gene expression. *Proc. Natl Acad. Sci. USA* **103**, 9993–9998 (2006).

Acknowledgements Cell sorting was performed with the assistance of A. Figueroa in the Core Facility of Department of Microbiology and Immunology at Columbia University. This work was supported by the Howard Hughes Medical Institute and by National Institutes of Health (NIH) grant R01 CA30488 (to S.P.G.).

Reviewer information *Nature* thanks A. Engelman, M. Lorincz, L. Naldini and M. Ott for their contribution to the peer review of this work.

Author contributions Y.Z., G.Z.W. and S.P.G. designed the experiments. Y.Z. and S.P.G. wrote the manuscript, with input from G.Z.W. and O.C. Y.Z. and G.Z.W. performed experiments in Fig. 2 and ChIP–qPCR experiments. All other experiments were performed by Y.Z.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0750-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0750-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.P.G.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Cell Lines. Mammalian cell lines, HeLa (CCL-2), NIH3T3 (CRL-1658), HEK293T (CRL-11268), COS-7 (CRL-1651) and the chicken cell line DF-1 (CRL-12203) were purchased from ATCC. All lines and their derivatives were maintained at 37 °C and 5% CO_2 in DMEM supplemented with 10% heat-inactivated fetal bovine serum, 2 mM glutamine, penicillin and streptomycin. MT-4 (#120) cells were obtained through the NIH AIDS Reagent Program and cultured at 37 °C and 5% CO_2 in RPMI 1640 supplemented with 10% heat-inactivated fetal bovine serum.

DNA construction. The replication-defective retroviral vectors pNCA-Luc (MLV vector that expresses firefly luciferase), pNCA-GFP (MLV vector that expresses GFP) and pSARM-Luc (MPMV vector that expresses firefly luciferase) have been described previously[19]. pRCAS-Luc (RSV vector that expresses luciferase) was constructed by replacing GFP in pRCAS-GFP with firefly luciferase. pNL4.3-Luc (HIV-1 vector that expresses firefly luciferase) was obtained from the NIH AIDS Reagent Program. The plasmid pCMV-intron expresses wild-type Gag and Pol from NB-tropic MLV. pMD2.G expresses the vesicular stomatitis virus (VSV) envelope glycoprotein. All integrase-deficient constructs (pRCAS-Luc(IN(D121A)), pSARM-Luc(IN(D127A)), pCMV-intron(IN(D184A)), pNL4.3-Luc(IN(D64A)) were created by site-directed mutagenesis.

pLvx-EF1-IRES-Neo was constructed by replacing the CMV promoter in pLvx-IRES-Neo (Clontech, 632181) with the EF1 promoter. The coding sequence of *MPP8* (wild type or with a W80A mutation) was cloned into pLvx-EF1-IRES-Neo to produce pLvx-EF1-IRES-Neo-MPP8 and pLvx-EF1-IRES-Neo-MPP8(W80A). The coding sequences of *NP220* with a silent mutation in the sgRNA targeting region (wild type), a DNA-binding-domain deletion NP220(ΔDB)) or with a zinc-finger-motif deletion (NP220(ΔZnF)) were also cloned into pLvx-EF1-IRES-Neo to produce pLvx-EF1-IRES-Neo-NP220, pLvx-EF1-IRES-Neo-NP220(ΔDB), pLvx-EF1-IRES-Neo-NP220(ΔZnF)).

MLV-Luc reporter vectors that contained deletions of or mutations in putative NP220-binding sites were constructed as follows: the MLV(U3) region in the 3′ LTR of pNCA-Luc was replaced by RSV(U3) to generate MLV-Luc-RSV(U3); MLV-Luc-U3(M1) was constructed by deleting −176 to −285 bp (relative to the first nucleotide of the R region) in the 3′ LTR U3 region; MLV-Luc-U3(M2) was constructed by deleting the first 3 putative NP220-binding sites (shown in Extended Data Fig. 6a); MLV-Luc-U3(M3) was constructed based on MLV-Luc-U3(M2) by further mutating the fourth and fifth putative NP220-binding sites (shown in Extended Data Fig. 6a) to TCTTCG and ACTTCT, respectively. We mutated rather than deleted the fourth and fifth binding sites, because they are located near the TATA box. All mutations or deletions were introduced into MLV-Luc reporter vectors by overlapping PCR.

pNL4.3-Luc-U3(M1), U3(M2) and U3(M3) were constructed by deleting the first three, five and all six putative NP220-binding sites (shown in Extended Data Fig. 7a), respectively. All mutations or deletions were introduced into MLV-Luc reporter vectors by overlapping PCR.

The ecotropic MLV infectious molecular clone (pNCS) has been deposited to Addgene (plasmid 17362). The amphotropic MLV infectious molecular clone (pNCA-Ampho) has been described previously[20]. The MLV HIV-1 NL4-3 infectious molecular clone (pNL4-3) was obtained from the NIH AIDS Reagent Program.

DNA transfection, virus packaging and infection. All DNA transfections were performed using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocol.

To package HIV-1- or MPMV-based VSV-G pseudotyped retroviruses, viral vectors (pNL4.3-Luc or pSARM-Luc, or their derivative mutants)—together with pMD2.G—were transfected into HEK293T cells. To package MLV-vector-based VSV-G pseudotyped viruses, viral vectors (pNCA-GFP or pNCA-Luc)—together with pCMV-intron (which expresses both MLV Gag and either wild-type Gag-Pol or Gag-Pol with mutant integrase) and pMD2.G—were transfected into HEK293T cells. To package RSV-based VSV-G pseudotyped retroviruses, viral vectors (pRCAS-Luc and its derivative mutants)—together with pMD2.G—were transfected into DF-1 cells. To package lentiviral-vector-based VSV-G pseudotyped viruses, viral vectors—together with pCMVdeltaR8.2 (which expresses HIV-1-Gag and -Gag–Pol) and pMD2.G—were transfected into HEK293T cells. Then, 40 h after transfection, supernatants were collected and filtered through a 45-μm membrane to produce virus preparations.

Unless otherwise indicated, viruses were diluted threefold with cell-culture medium containing 20 mM HEPES (pH 7.5) and 4 μg/ml polybrene. The cell-culture medium was changed 3 h after infection.

Luciferase assay. Luciferase activity was assayed 40 h after infection, using the Luciferase Assay System (Promega).

Reverse transcription. Reverse transcription was performed using a High-Capacity cDNA Reverse Transcription Kit (ThermoFisher Scientific).

Quantitative PCR. Quantitative PCR (qPCR) was performed in an ABI 7500 Fast Real-Time PCR System using FastStart Universal SYBR Green Master (Rox). The PCR cycle program was: (1) 50 °C, 2 min, 1 cycle; (2) 95 °C, 10 min, 1 cycle; (3) from 95 °C, 15 s to 60 °C, 30 s to 72 °C, 30 s, 40 cycles; (4) 72 °C, 10 min, 1 cycle. The primers used for qPCR are listed in Supplementary Table 2.

ChIP. In brief, 2×10^6 cells were seeded in 10-cm dishes and infected with VSV-G pseudotyped, integrase-deficient or integrase-proficient MLV-Luc viruses for two days. The virus used for infection was pretreated with 5 U ml^{-1} DNase (Promega, M6101) supplemented with 10 mM $MgCl_2$ at 37 °C for 1 h to remove any residual plasmid DNA. Cells were crosslinked in 1% formaldehyde for 10 min, quenched in 0.125 M glycine for 5 min and lysed in 1 ml of ChIP lysis buffer (50 mM Tris-HCl pH 8.0, 1% SDS, 10 mM EDTA, complemented with protease inhibitor cocktail). Cell lysates were sonicated using a Branson 450 Digital Sonifier (12% power amplitude, sonication for 30 s (eight times) on ice with 60 s between each sonication) to produce an average chromatin fragment size of 200–800 base pairs and lysates were subsequently centrifuged at 13,000 r.p.m. at 4 °C for 20 min. The supernatant of approximately 50 μg sonicated chromatin was then immunoprecipitated overnight using 5 μg of the specific antibodies in ChIP dilution buffer (10 mM Tris-HCl pH 8.0, 1% Triton X-100, 0.1% SDS, 150 mM NaCl, 2 mM EDTA). The next day, 25 μl of Dynabeads (12.5 μl protein A and 12.5 μl protein G) was added and the mixture was incubated for an additional 2 h. The beads were washed twice each in ChIP low-salt buffer (20 mM Tris-HCl pH 8.0, 1% Triton X-100, 0.1% SDS, 150 mM NaCl, 2 mM EDTA), ChIP high-salt buffer (20 mM Tris-HCl pH 8.0, 1% Triton X-100, 0.1% SDS, 500 mM NaCl, 2 mM EDTA), ChIP LiCl buffer (10 mM Tris-HCl pH 8.0, 1% NP-40, 250 mM LiCl, 1 mM EDTA) and TE buffer (10 mM Tris-HCl pH8.0, 1 mM EDTA). Protein–DNA complexes were eluted from beads in 200 μl elution buffer (TE buffer containing 1% SDS, 100 mM NaCl, 5 mM DTT), treated with RNase A (1 μg per elution, 37 °C, 1 h) and proteinase K (15 μg per elution, 37 °C, 2 h), reverse crosslinked (65 °C overnight) and purified using the QIAquick PCR purification kit according to the manufacturer's instructions (Qiagen). qPCRs were performed with the indicated primers (see Supplementary Table 2). For all ChIP assays, ChIP experiments with histone H3 and control rabbit IgG antibodies were included as positive control and negative control, respectively. qPCR data from each ChIP assay with specific antibodies were calculated as a percentage relative to input DNA.

CRISPR-mediated gene knockout. Four sgRNAs per gene were selected from the human CRISPR knockout pooled library (Brunello, Addgene 73178) and cloned into LentiCRISPRv2GFP (for *MPP8* knockout) or lentiCRISPR v.2 (for all the other genes). HeLa cells were transfected with a pool of four plasmids using Lipofectamine 2000 for two days. For *MPP8* knockout, the brightest 1% GFP^+ cells were sorted by fluorescence-activated cell sorting (FACS). For the knockout of other genes, the transfected cells were selected in 1 μg ml^{-1} puromycin for two weeks. Single cells from the resulting pool cells were seeded in a 96-well plate and specific gene knockout clones were screened by western blotting using specific antibodies.

To rescue the expression of MPP8 in *MPP8* knockout cells, *MPP8* knockout cells were transduced with pLvx-EF1-IRES-Neo-MPP8 (wild type or with the indicated mutation) and selected in 800 μg ml^{-1} G418 for two weeks. To rescue the expression of NP220 in *NP220* knockout cells, *NP220* knockout cells were transduced with pLvx-EF1-IRES-Neo-NP220 (wild type or with the indicated deletion) and selected in 800 μg ml^{-1} G418 for two weeks.

To knockout *NP220* in MT-4 cells, four target sequences were selected from the human CRISPR knockout pooled library (Brunello, Addgene 73178). To knockout *Np220* in NIH3T3 cells, four targets were selected from the mouse CRISPR knockout pooled library (Brie, Addgene 73633). For each knockout, a mix of four CRISPR RNAs (crRNAs) was synthesized by Integrated DNA Technologies (IDT). The crRNAs were first annealed with ATTO550 tagged Alt-R CRISPR–Cas9 trans-activating crRNA (tracrRNA) (IDT 1075927) and then mixed with *Streptococcus pyrogenes* Cas9 nuclease (IDT 1081058) at room temperature for 20 min to form ribonucleoprotein (RNP) complexes. The final concentrations of the crRNA:tracrRNA duplex and Cas9 nuclease were 24 μM and 20.8 μM. For NIH3T3 cells, 5 μl of the RNP was mixed with 3.5×10^5 NIH3T3 cells and RNP transfection was performed using a 4D-Nucleofector System (Lonza) with program EN-158. For MT-4 cells, 5 μl of the RNP were mixed with 10^6 MT-4 cells and RNP transfection was performed using a 4D-Nucleofector System (Lonza) with program CA-137. Subesequently, 24 h after transfection, 1% brightest ATTO 550 cells were sorted by FACS and expanded. The knockout efficiency was confirmed by western blotting using a NP220-specific antibody.

siRNA transfection. All siRNAs were obtained from Dharmacon (SMARTpool: ON-TARGETplus siRNA). Targeted gene and catalogue numbers (indicated in brackets) are as follows: *HDAC1* (L-003493-00), *HDAC2* (L-003495-02), *HDAC3* (L-003496-00), *HDAC4* (L-003497-00), *HDAC5* (L-003498-00), *HDAC6* (L-003499-00), *HDAC7* (L-009330-00), *HDAC8* (L-003500-00), *HDAC9* (L-005241-00), *HDAC10* (L-004072-00), *HDAC11* (L-004258-00), *NP220* (human, L-013715-02), *Np220* (mouse, L-047870-01), *MPP8* (L-021680-02), *TASOR* (L-020306-02), *PPHLN1* (L-021306-02), *GGT1* (L-005884-01), *CISD2* (L-032593-02),

© 2018 Springer Nature Limited. All rights reserved.

ZNT1 (L-007522-01), *DUOXA2* (L-032210-02), *TESC* (L-020896-01), *GP5* (L-012671-00), *SETDB1* (L-020070-00), *SETDB2* (L-014751-00), *SUV39H1* (L-009604-00), *SUV39H2* (L-008512-00), *EHMT1* (L-007065-02), *EHMT2* (L-006937-00), *EZH2* (L-004218-00) and non-targeting control siRNA (NT, D-001810-10).

For siRNA transfection, 10^5 HeLa cells were seeded in six-well plates. After 24 h, siRNAs were transfected into the cells using Lipofectamine RNAiMax (Invitrogen) according to the manufacturer's protocol. After another 24 h, the same siRNA transfection was performed for the second time. Then, 6 h after second transfection, the siRNA transfected cells were infected with virus for further experiments.

Co-immunoprecipitation and immunoblotting. For co-immunoprecipitation, approximately 5×10^6 HeLa cells were lysed in 1 ml Pierce IP Lysis Buffer for 10 min and the lysate was clarified by centrifugation at 4 °C for 15 min at 12,000 r.p.m. For nuclease treatment, cell lysates were treated with benzonase (250 U ml^{-1} supplemented with 10 mM $MgCl_2$), DNase (5 U ml^{-1} supplemented with 10 mM $MgCl_2$) or RNase A (5 μg ml^{-1}). First, 40 μl Dynabeads (20 μl protein A and 20 μl protein G beads) were mixed with 1 μg of the specific antibody for 10 min at room temperature and then washed twice with TBST. Antibody-coated Dynabeads were incubated with 800 μl cell lysate at 4 °C for 4 h. The beads were then washed three times with TBST. The bound proteins on the beads were eluted with 40 μl 1× SDS sample buffer and subjected to SDS–PAGE and western blot analysis.

For immunoblotting, cells were lysed in RIPA Lysis and Extraction Buffer for 10 min. The lysate was clarified by centrifugation at 4 °C for 15 min at 12,000 r.p.m. The samples were heated at 95 °C in SDS sample buffer and resolved by SDS–PAGE electrophoresis, transferred to a PVDF membrane and probed with specific antibodies by western blot.

Genome-wide CRISPR–Cas9 screen. The human CRISPR knockout pooled library (Brunello, two-vector system) was obtained from Addgene and the screen was performed broadly as described previously[21]. The CRISPR sgRNA library virus was packaged by transfecting HEK293T cells with library DNA, a HIV-1-Gag–Pol-expressing plasmid, pCMVΔR8.2 and pMD2.G. HeLa cells were transduced with lentiCas9-Blast virus and two days after transduction, cells were selected in 5 μg ml^{-1} blasticidin for two weeks to get pooled HeLa Cas9 cells. 10^8 HeLa Cas9 cells were transduced with CRISPR sgRNA library viruses at a multiplicity of infection (MOI) of approximately 0.3. Two days after transduction, cells were selected in medium containing 1 μg ml^{-1} puromycin for two weeks to get the collection of pooled HeLa knockout cells and cells were cultured in medium containing 5 μg ml^{-1} blasticidin and 1 μg ml^{-1} puromycin during the whole process of screening. Then, 2×10^7 pooled knockout HeLa cells were infected with the integrase-deficient (IN(D184A)) MLV-GFP virus (threefold dilution) and the 5% brightest GFP cells were sorted by FACS. The sorted cells were expanded for two weeks and then the above infection/sorting procedure was performed for the second time. Genomic DNA was extracted from the resulting selected cells and the control cells (transduced with the sgRNA library, cultured in parallel but without infection/sorting). The abundances of sgRNAs in the control cells and the sorted cells were analysed by PCR followed by next generation sequencing. The PCR to amplify the sgRNA was performed as described in step 32–33 of the previously published protocol[21], except that the reverse primer for the control sample is: 5′-CAAGCAGAAGACGGCATACGAGATACTGTATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATTCTACTATTCTTTCCCCTGCACTGT-3′ and the reverse primer for the sorted sample is: 5′-CAAGCAGAAGACGGCATACGAGATAGGTCGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGATTCTACTATTCTTTCCCCTGCACTGT-3′. The PCR products were purified by Zymo-Spin V with Reservoir and the samples were deep-sequenced on the Illumina Miseq. The sgRNA counts and abundance were analysed as described[21]. The degree of sgRNA enrichment and gene hit rank was analysed by software HiTSelect[22].

Bacterial protein purification. The gene fragment encoding the NP220 DNA-binding domain (NP220(DB)), corresponding to amino acid residues 1240–1478 of NP220) was cloned into the pGEX-5X-3 vector. The protein was overexpressed in *Escherichia coli* BL21 (DE3) Star strain (NEB). The cells were induced with 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 4 h at 37 °C. The cells were then collected and resuspended in lysis buffer, containing 50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.05% NP-40 and lysed with 0.25 mg ml^{-1} lysozeme for1 h followed by sonication. Cell lysates were centrifuged for 30 min at 4 °C before incubating with glutathione sepharose beads at 4 °C for 2 h. The beads were washed five times with PBS and proteins were eluted by elution buffer containing 20 mM Tris-HCl (pH 8.0), 100 mM NaCl, 2 mM $CaCl_2$ and 10 μg ml^{-1} factor Xa protease (NEB).

Electrophoretic mobility shift assays. The DNA probe biotin–DNA84 was 5′-end labelled with biotin. Electrophoretic mobility shift assays (EMSA) were performed using a LightShift Chemiluminescent EMSA Kit (Thermo Fisher). In brief, 20 fmol biotin–DNA84 was incubated with 800 ng bacterially expressed NP220(DB) protein at room temperature in a total volume of 20 μl. For DNA competition experiments, unlabelled DNA (DNA84 or DNA72) was added at the same time as the probe. Binding reactions were analysed by electrophoresis on 5% native polyacrylamide gels. The sequence of DNA84 is: AGGATATCTGTGGTAAGCAGTTCCTGCCCCGGCTCAGGGCCA AGAACAGATGGTCCCCAGATGCGGTCCAGCCCTCAGCAGTTT (NP220 binding sites are underlined), and the sequence of DNA72 is: AGGATATCTGTGGTAAGCAGTTCCTGCTCAGGGCCAAGAACAGATGGTATGCGGTCCAGCCCTCAGCAGTTT. All DNA duplexes were ordered from IDT.

MLV replication and reverse transcriptase assay. Ecotropic MLV and amphotropic MLV viruses were produced by transfecting HEK293T cells with pNCS and pNCA-Ampho DNA, respectively, using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocol.

For ecotropic MLV replication in NIH3T3 cells, *Np220* wild-type or *Np220* knockout cells (10^4 cells per well) were seeded in six-well plates and infected with ecotropic MLV at a low MOI. Culture medium was changed at 3 h after infection. Culture supernatants (50 μl) were taken every day for five days after infection and assayed for reverse transcriptase activity to monitor the production of progeny virus.

For amphotropic MLV replication in HeLa cells, *NP220* wild-type or knockout cells (3×10^4 cells per well) were seeded in six-well plates and infected with amphotropic MLV at a low MOI. Culture medium was changed at 3 h after infection. Culture supernatants (50 μl) were taken every two days for twelve days after infection. Cells were split 1:20 at six days after infection. The culture medium was assayed for reverse transcriptase activity to monitor the production of progeny virus.

For the reverse transcriptase assay, 5 μl of culture supernatant was incubated with 20 μl hot/cold mix at room temperature for 40 min and then 5 μl of the reaction mix was spotted on DEAE paper. The DEAE paper was washed with 2× SSC buffer for 20 min twice, dried under a heat lamp, exposed and visualized by phosphor imaging (GE). The formula for the buffers are as following: hot/cold mix (1 ml cold mix, 72 μl hot mix, 2 μl $MnCl_2$ (0.5 M), 1 μl ^{32}P dTTP, add all reagents in the indicated order); cold mix (60 mM Tris-HCl pH 8.3, 75 mM NaCl, 0.06% NP-40); hot mix (7.6 mg ml^{-1} oligo(dT), 16.6 μM dTTP, 166 μg μl^{-1} poly(A), 0.5 M DTT); oligo(dT) (GE healthcare, 27-7858-2); poly(A) (GE healthcare, 27-4110).

HIV-1 replication. HIV-1 viruses were produced by transfecting HEK293T cells with pNL4.3 using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocol. For HIV-1 replication in MT-4 cells, 10^6 *NP220* wild-type or knockout cells were transduced with HIV-1 virus (1 ng capsid protein p24) by spin infection (3,500 r.p.m. at room temperature for 2 h) and then cultured at 37 °C. After 3 h of infection, cells were washed twice with PBS and then cultured in 1 ml medium in a 24-well plate. Every two days for twelve days, aliquots of the culture supernatants (50 μl) were taken for p24 measurement and half of the cells and medium (500 μl) were transferred to a new well containing 500 μl fresh medium. p24 levels were determined by enzyme-linked immunosorbent assay (ELISA) using the HIV1 p24 ELISA kit (Abcam, ab218268), to monitor the production of progeny virus.

Statistical analysis. Sample sizes are provided in the figure legends. Statistical significance was determined by a two-tailed Student's *t*-test. The experiments were not randomized. The investigators were not blinded to allocation during experiments or outcome assessment.

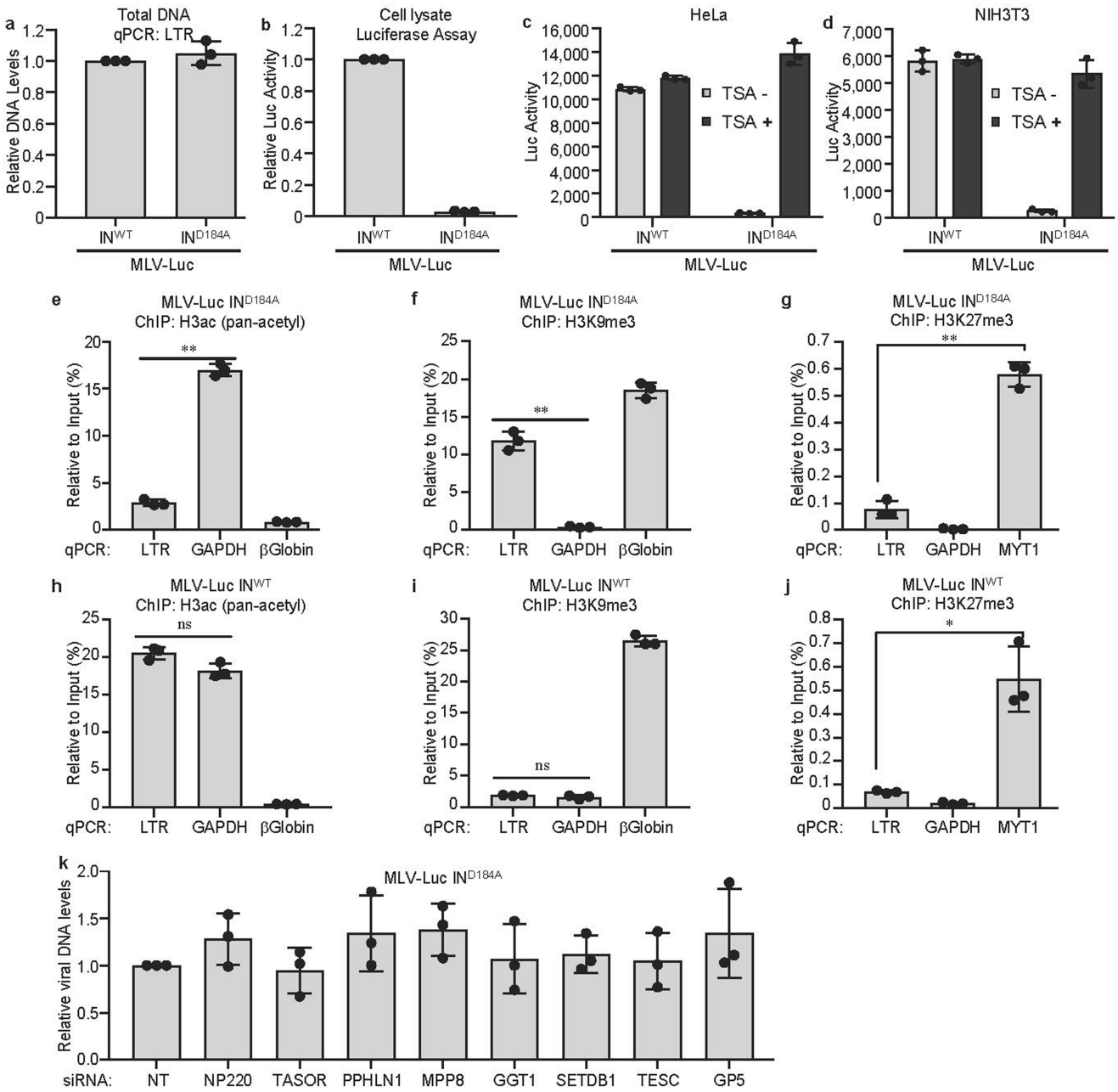
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data that support the findings of this study are available from the corresponding author upon request. Source Data is available for the Figs 3a–g, 4c–h, 5b, d and Extended Data Figs. 1e–j, 3a–e, 4a–d, 6f, g. For gel source data, see Supplementary Fig. 1.

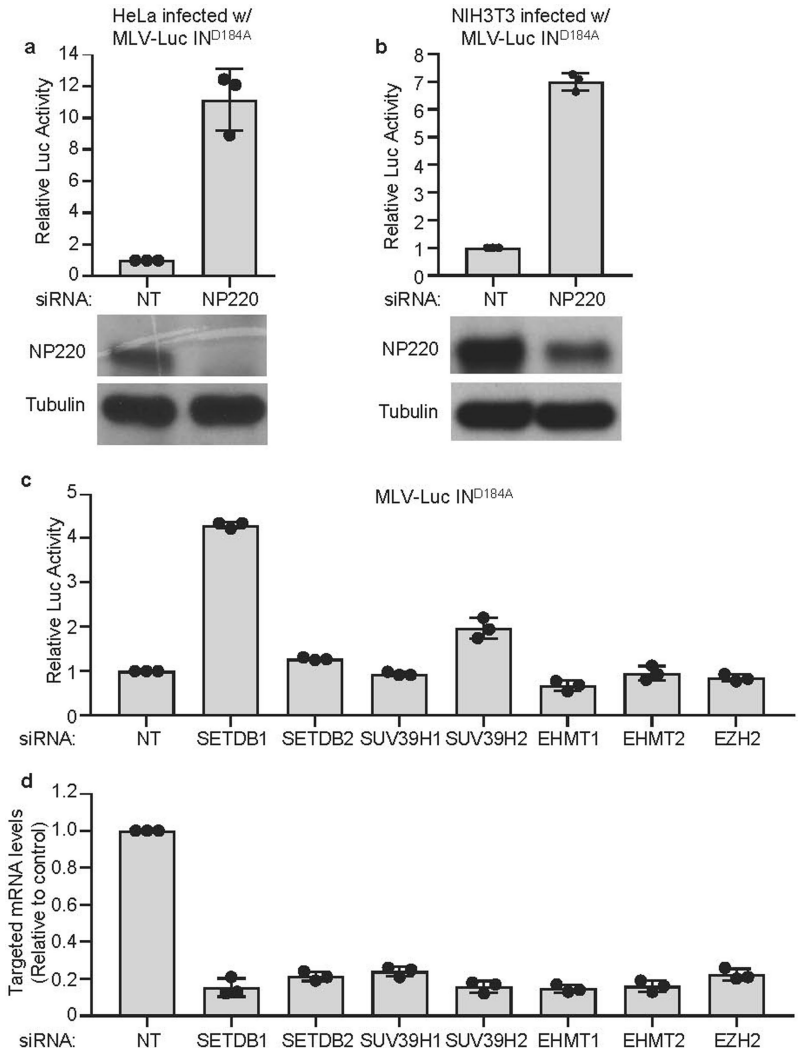
19. Wang, G. Z. & Goff, S. P. Postentry restriction of Mason–Pfizer monkey virus in mouse cells. *J. Virol.* **89**, 2813–2819 (2015).
20. Wolf, D. & Goff, S. P. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature* **458**, 1201–1204 (2009).
21. Joung, J. et al. Genome-scale CRISPR–Cas9 knockout and transcriptional activation screening. *Nat. Protoc.* **12**, 828–863 (2017).
22. Diaz, A. A., Qin, H., Ramalho-Santos, M. & Song, J. S. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic Acids Res.* **43**, e16 (2015).

© 2018 Springer Nature Limited. All rights reserved.



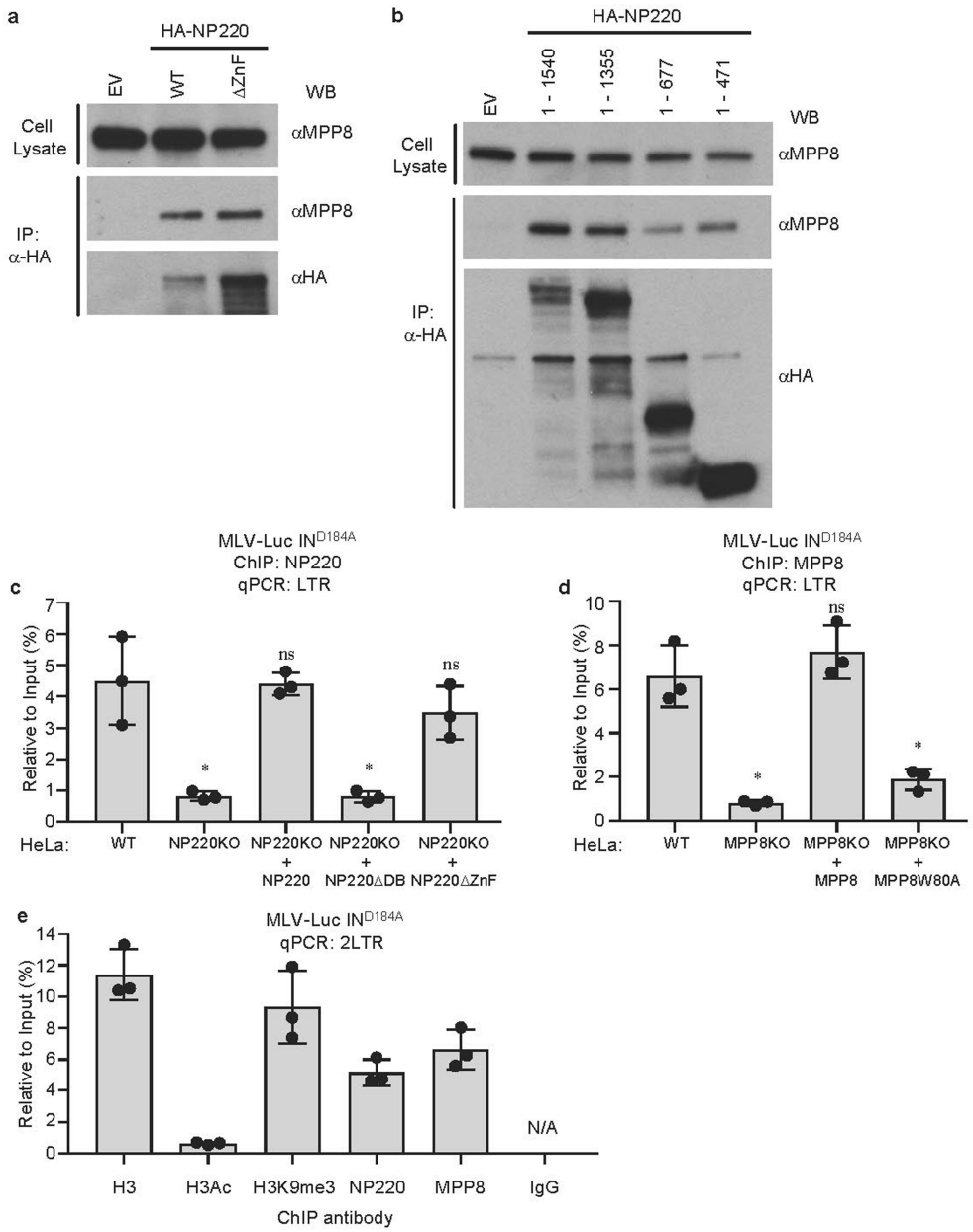
Extended Data Fig. 1 | Silencing of unintegrated, but not integrated, retroviral DNA. a, **b**, Silencing of unintegrated retroviral DNA. HeLa cells were infected with VSV-G pseudotyped, integrase-proficient (IN(WT)) or integrase-deficient (IN(D184A)) MLV-Luc viruses. Total viral DNA levels (**a**) and luciferase activities (**b**) were measured 40 h after infection. The viral DNA concentration and luciferase activity of IN(WT) MLV-Luc were set to 1. Data are mean ± s.d.; $n = 3$ independent experiments. **c**, **d**, Silencing of unintegrated retroviral DNA is dependent on histone deacetylation. HeLa (**c**) and NIH3T3 (**d**) cells were infected with VSV-G pseudotyped, integrase-proficient (IN(WT)) or integrase-deficient (IN(D184A)) MLV-Luc viruses and treated with DMSO (TSA−) or 1 μM HDAC inhibitor trichostatin A (TSA+). Luciferase activities were measured 40 h after infection. Data are mean ± s.d.; $n = 3$ independent experiments. **e**–**j**, Repressive epigenetic marks are present on unintegrated retroviral DNA, whereas active epigenetic marks are present on integrated retroviral DNA. HeLa cells were infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) (**e**–**g**) or integrase-proficient (IN(WT)) (**h**–**j**) MLV-Luc viruses. At 40 h after infection, ChIP—using antibodies against pan-acetyl H3, H3K9me3 or H3K27me3—followed by qPCR using indicated primers, was performed to assess H3ac (**e**, **h**), H3K9me3 (**f**, **i**) and H3K27me3 (**g**, **j**) modifications across the LTR of unintegrated and integrated MLV-Luc DNA. qPCR data from each ChIP were calculated as the percentage of input DNA. Data are mean ± s.d.; $n = 3$ independent experiments. ns, $P > 0.05$; *$P < 0.05$; **$P < 0.01$. P values are from paired two-sided Student's t-tests. Exact P values are included in the Source Data associated with this figure. **k**, Knockdown of indicated genes has no effect on viral DNA levels. HeLa cells were first transfected with the indicated siRNA and then infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus. Viral DNA levels were measured 40 h after infection. DNA levels in HeLa cells transfected with a non-targeting (NT) control siRNA was set to 1. Data are mean ± s.d.; $n = 3$ independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



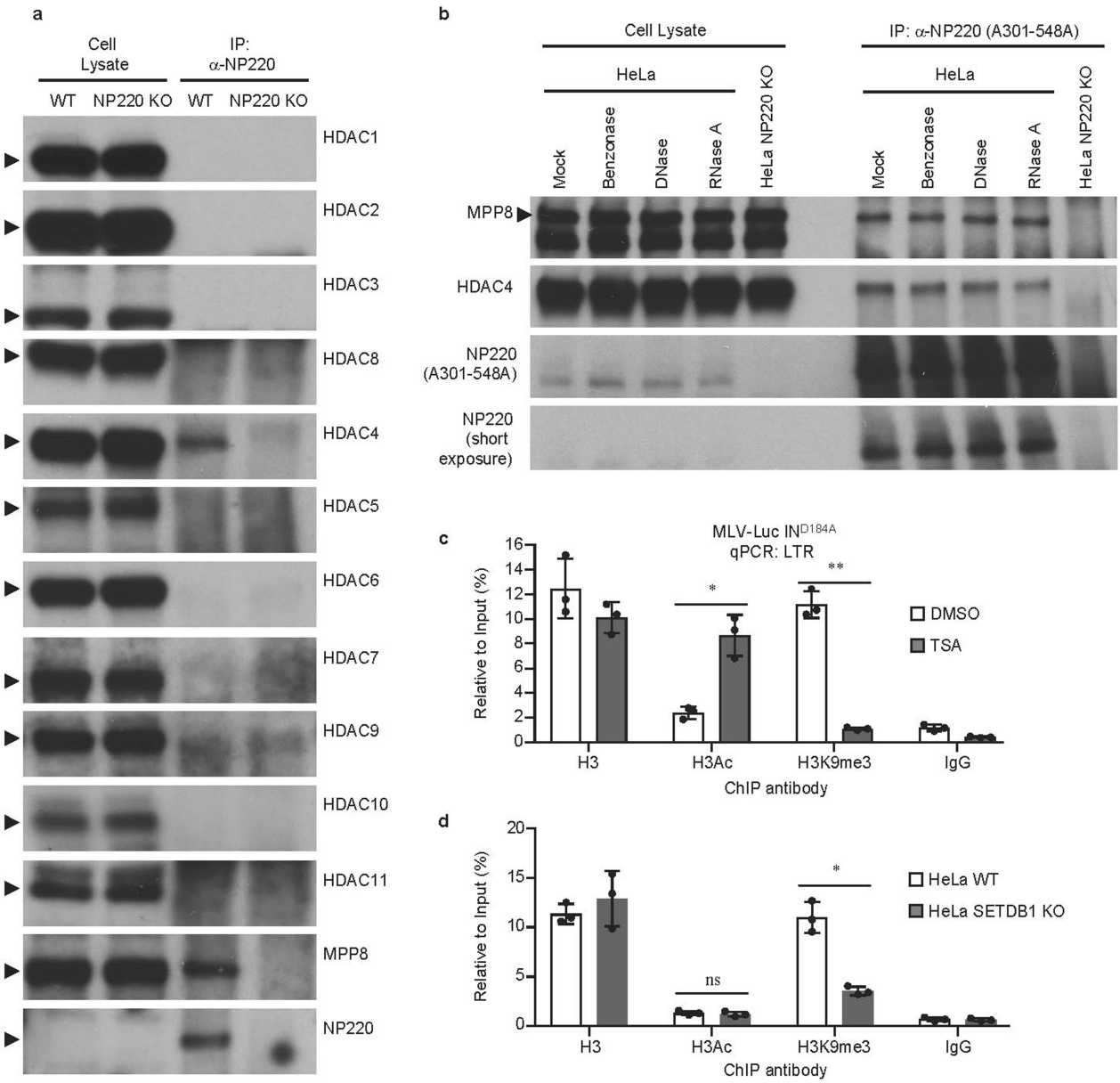
Extended Data Fig. 2 | NP220 and SETDB1 are required for the silencing of unintegrated MLV DNA. a, **b**, HeLa (**a**) or NIH3T3 (**b**) cells were transfected with indicated siRNAs and then infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus. Luciferase activities were measured 40 h after infection and luciferase activity in non-targeting (NT) control siRNA-transfected cells was set to 1 (top). Data are mean ± s.d.; $n = 3$ independent experiments. The expression of NP220 was determined by western blot (bottom). **c**, HeLa cells were first transfected with the indicated siRNAs targeting histone methyltransferases and then infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus. Luciferase activities were measured 40 h after infection and luciferase activity in non-targeting (NT) control siRNA-transfected HeLa cells was set to 1. Data presented are mean ± s.d.; $n = 3$ independent experiments. **d**, Knockdown efficiency of siRNAs used in **c**. HeLa cells were transfected with the indicated siRNAs targeting histone methyltransferases and mRNA levels of siRNA-targeted genes were measured by qPCR with reverse transcription (RT–qPCR). mRNA levels in non-targeting (NT) control siRNA-transfected HeLa cells was set to 1. Data are mean ± s.d.; $n = 3$ independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



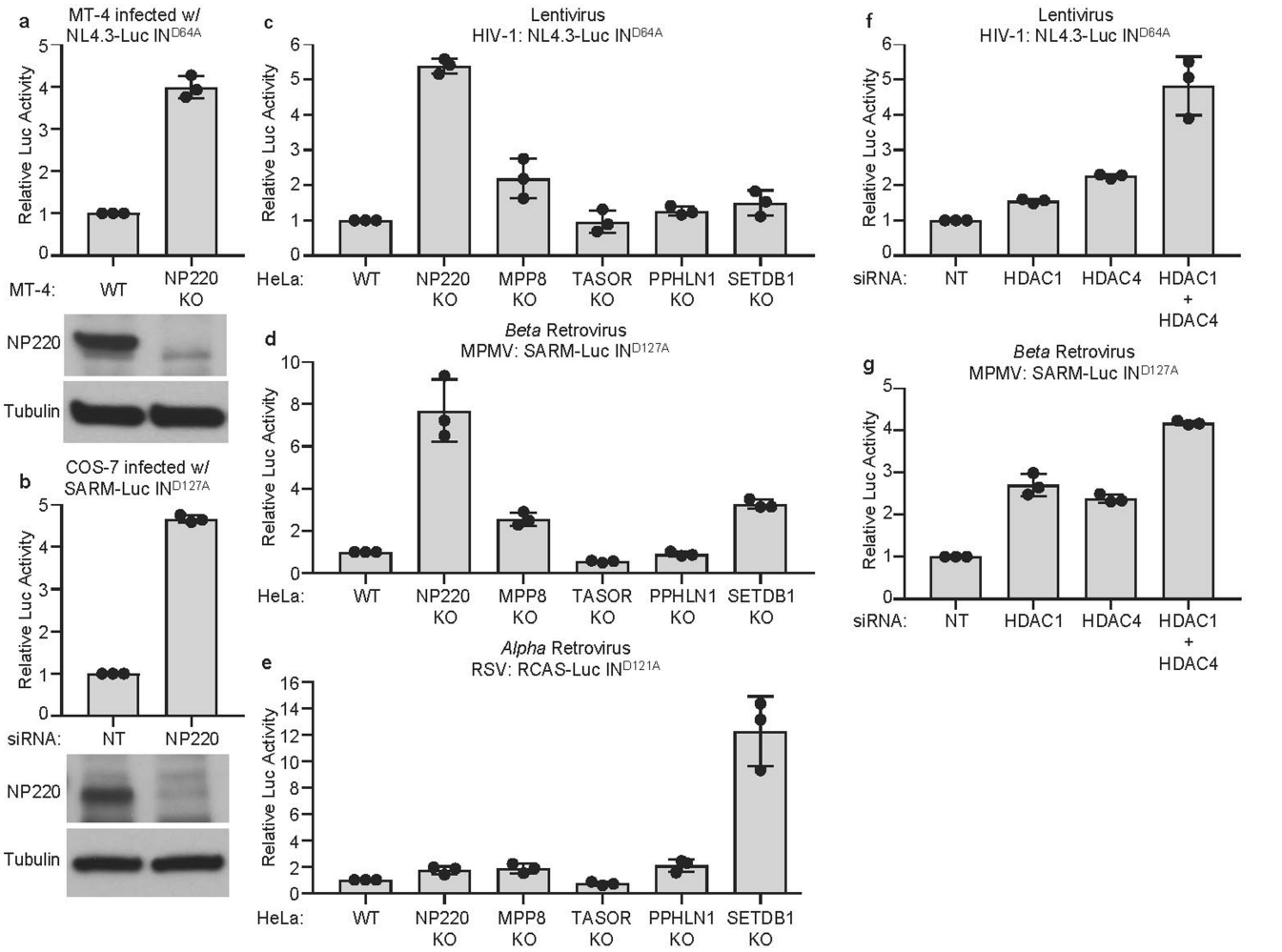
Extended Data Fig. 3 | NP220 recruits the HUSH complex and SETDB1 to silence unintegrated MLV DNA. a, **b**, HA-tagged NP220, NP220 with zinc finger deletion (ΔZnF) (**a**) or indicated fragments of NP220 (**b**) were introduced into *NP220* knockout HeLa cells and then immunoprecipitated using an HA antibody. Co-immunoprecipitated MPP8 was analysed by western blot. Images are representative of two independent experiments with similar results. **c**, Parental wild-type (WT) HeLa cells, *NP220* knockout HeLa cells and *NP220* knockout HeLa cells that were reconstituted with indicated variants of NP220 were subsequently infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus. At 40 h after infection, ChIP was performed to assess the association of NP220 across the LTR of unintegrated MLV-Luc DNA. qPCR data from each ChIP were calculated as the percentage of input DNA. Data are mean ± s.d.; $n = 3$ independent experiments. ns, $P > 0.05$; $^*P < 0.05$. *P* values are from paired two-sided Student's *t*-tests. Exact *P* values are included in the Source Data associated with this figure. **d**, Parental HeLa cells, *MPP8* knockout HeLa cells and *MPP8* knockout HeLa cells that were reconstituted with indicated variants of MPP8 were infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus. At 40 h after infection, ChIP was performed to assess the association of MPP8 with the LTR of unintegrated MLV-Luc DNA. qPCR data from each ChIP were calculated as the percentage of input DNA. Data are mean ± s.d.; $n = 3$ independent experiments. ns, $P > 0.05$; $^*P < 0.05$. *P* values are from paired two-sided Student's *t*-tests. Exact *P* values are included in the Source Data associated with this figure. **e**, HeLa cells were infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus. At 40 h after infection, ChIP was performed using the indicated antibodies followed by qPCR using primers targeting 2LTR circles. qPCR data from each ChIP were calculated as the percentage of input DNA. Data are mean ± s.d.; $n = 3$ independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Interaction between NP220 and HDACs. **a**, Screen for interactions between NP220 and HDACs. Endogenous NP220 were immunoprecipitated (IP) from indicated HeLa cell lines and the indicated co-immunoprecipitated HDAC proteins and MPP8 were analysed by western blot using specific antibodies. MPP8 serves as a positive control. Images are representative of two independent experiments with similar results. **b**, NP220–MPP8 or NP220–HDAC4 interactions are independent of DNA or RNA. Cell lysates from indicated HeLa cell lines were treated with benzonase, DNase or RNase A, endogenous NP220 were immunoprecipitated and co-immunoprecipitating proteins were analysed by western blot using specific antibodies. Images are representative of two independent experiments with similar results. **c**, **d**, Interrelationship of histone deacetylation and H3K9 trimethylation on unintegrated viral DNA. **c**, HeLa cells were infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus and treated with DMSO (TSA−) or 1 μM HDAC inhibitor trichostatin A (TSA+). **d**, Parental HeLa cells and *SETDB1* knockout HeLa cells (*SETDB1* KO) were infected with VSV-G pseudotyped, integrase-deficient (IN(D184A)) MLV-Luc virus. At 40 h after infection, ChIP was performed using the indicated antibodies followed by qPCR using primers targeting LTR. qPCR data from each ChIP were calculated as the percentage of input DNA. Data are mean ± s.d.; $n = 3$ independent experiments. $^{*}P < 0.05$; $^{**}P < 0.01$. *P* values are from paired two-sided Student's *t*-tests. Exact *P* values are included in the Source Data associated with this figure.

© 2018 Springer Nature Limited. All rights reserved.

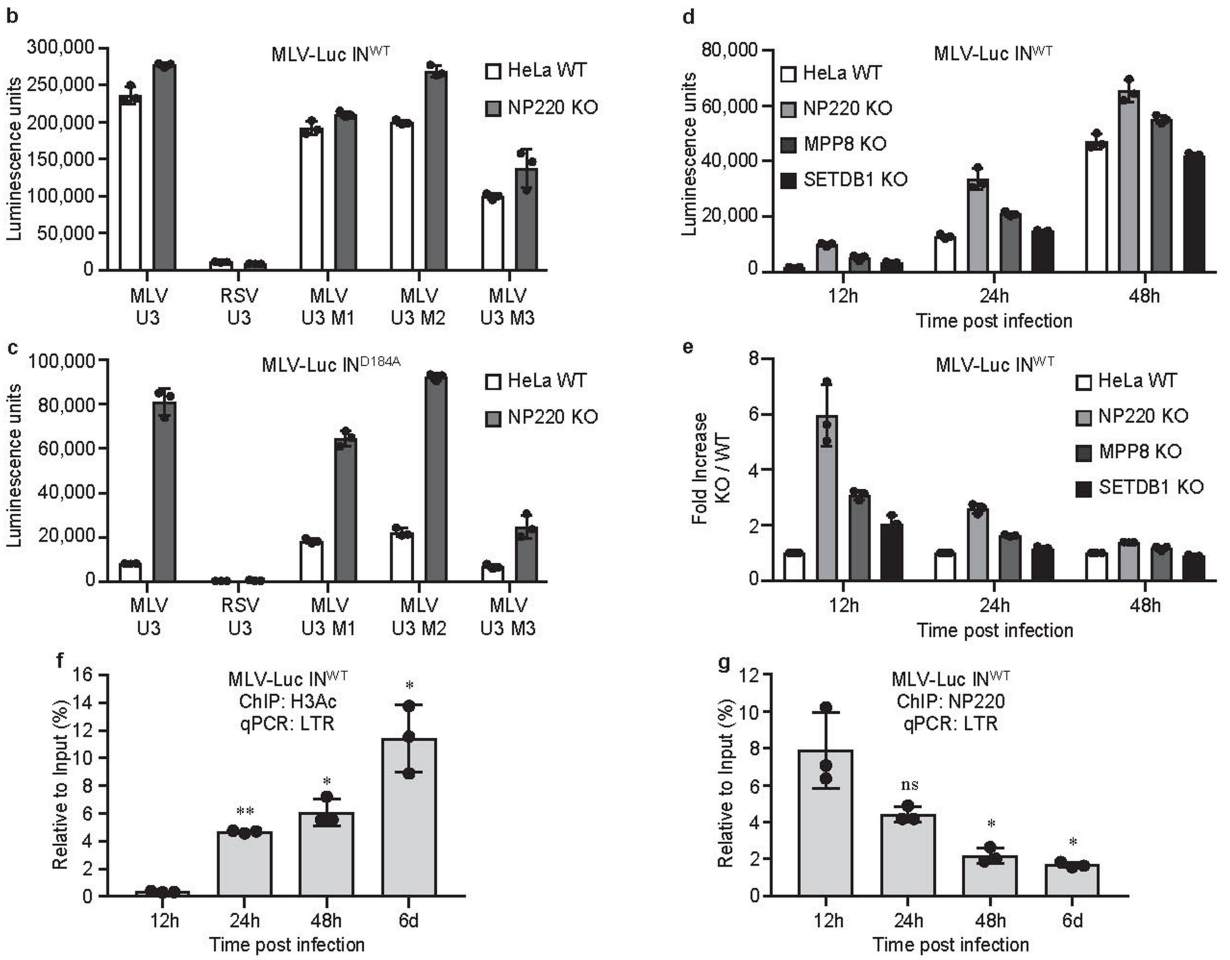


Extended Data Fig. 5 | NP220 mediates silencing of unintegrated retroviral DNA from HIV-1 and MPMV, but not RSV. a, Parental MT-4 cells and *NP220* knockout MT-4 cell line (*NP220* KO) were infected with VSV-G pseudotyped, integrase-deficient (IN(D64A)) HIV-1 vector NL4.3-Luc. Luciferase activities were measured 40 h after infection and luciferase activity in parental (wild-type) MT-4 cells was set to 1 (top). Data are as mean ± s.d.; $n = 3$ independent experiments. The expression of NP220 was determined by western blot (bottom). **b**, COS-7 cells were first transfected with indicated siRNAs and then infected with VSV-G pseudotyped, integrase-deficient (IN(D127A)) MPMV vector SARM-Luc. Luciferase activities were measured 40 h after infection and luciferase activity in non-targeting (NT) control siRNA-transfected cells was set to 1 (top). Data are mean ± s.d.; $n = 3$ independent experiments. The expression of NP220 was determined by western blot (bottom). **c**–**e**, Indicated HeLa cell lines were infected with VSV-G pseudotyped, integrase-deficient HIV-1 vector NL4.3-Luc (**c**), MPMV vector SARM-Luc (**d**) or RSV vector RCAS-Luc (**e**). Luciferase activities were measured 40 h after infection. Luciferase activity in parental HeLa cells was set to 1. Data are mean ± s.d.; $n = 3$ independent experiments. **f**, **g**, HDAC1 and HDAC4 are involved in silencing of unintegrated HIV-1 and MPMV DNA. HeLa cells were first transfected with the indicated siRNAs and then infected with VSV-G pseudotyped, integrase-deficient (IN(D64A)) HIV-1 vector NL4.3-Luc (**f**) or VSV-G pseudotyped, integrase-deficient (IN(D127A)) MPMV vector SARM-Luc (**g**). Luciferase activities were measured 40 h after infection and luciferase activities in HeLa cells transfected with control non-targeting (NT) siRNA were set to 1. Data are mean ± s.d.; $n = 3$ independent experiments.

© 2018 Springer Nature Limited. All rights reserved.

a >>MLV U3

ACCTGTAGGTTTGGCAAGCTAGCTTAAGTAACGCCATTTTGCAAGGCATGGAAAAATACATAACTGAGAATAGAGAAGTTCAGATCAAGGTCAGGAACAGATGGAACAGCTG
AATATGGGCCAAACAGGATATCTGTGGTAAGCAGTTCCT**GCCCCG**GCTCAGGGCCAAGAACAGATGGAACAGCTGAATATGGGCCAAAC*AGGATATCTGTGGTAAGCAGTTC*
*CT**GCCCCG**GCTCAGGGCCAAGAACAGATGGT**CCCCAG**ATGCGGTCCAGCCCTCAGCAGTTT*CTAGAGAACCATCAGATGTTTCCAGGGTGCCCCAAGGACCTGAAATGACCC
TGTGCCTTATTTGAACTAACCAATCAGTTCGCTTCTCGCTTCTGTTCGCGCGCTTCTGC**TCCCCG**AGCTCAATAAAAGAGCCCACA**ACCCCT**CACTCGGC



Extended Data Fig. 6 | The binding specificity of NP220 bound to unintegrated MLV DNA. a, The sequence of MLV U3 region. Putative NP220-binding sites are indicated in red. The sequence of the 84-nucleotide probe used for EMSA is italicized and underlined. **b**, **c**, Parental or *NP220* knockout HeLa cells were infected with VSV-G pseudotyped, integrase-proficient (IN(WT)) (**b**) or integrase-deficient (IN(D184A)) (**c**) MLV-Luc viruses bearing the indicated deletions or mutations in the U3 region. Luciferase activities were measured 40 h after infection. Data are mean ± s.d.; $n = 3$ independent experiments. **d**, **e**, Indicated HeLa cell lines were infected with VSV-G pseudotyped, integrase-proficient (IN(WT)) MLV-Luc virus. **d**, At the indicated times after infection, luciferase activities were measured. **e**, Fold increase (KO/WT) was calculated as the ratio of the luciferase activities of knockout cells compared to the luciferase activities of wild-type cells. Data are mean ± s.d.; $n = 3$ independent experiments. **f**, **g**, The dynamics of H3Ac deposition and NP220 association on viral DNA during the course of MLV infection. HeLa cells were infected with VSV-G pseudotyped, integrase-proficient (IN(WT)) MLV-Luc virus. At the indicated time points after infection, ChIP was performed using antibodies against H3Ac (**f**) and NP220 (**g**) followed by qPCR using primers targeting the LTR, to monitor the association of H3Ac (**f**) and NP220 (**g**) with viral DNA. qPCR data from each ChIP were calculated as the percentage of input DNA. Data are mean ± s.d.; $n = 3$ independent experiments. ns, $P > 0.05$; $^*P < 0.05$; $^{**}P < 0.01$. *P* values are from paired two-sided Student's *t*-tests. Exact *P* values are included in the Source Data associated with this figure.

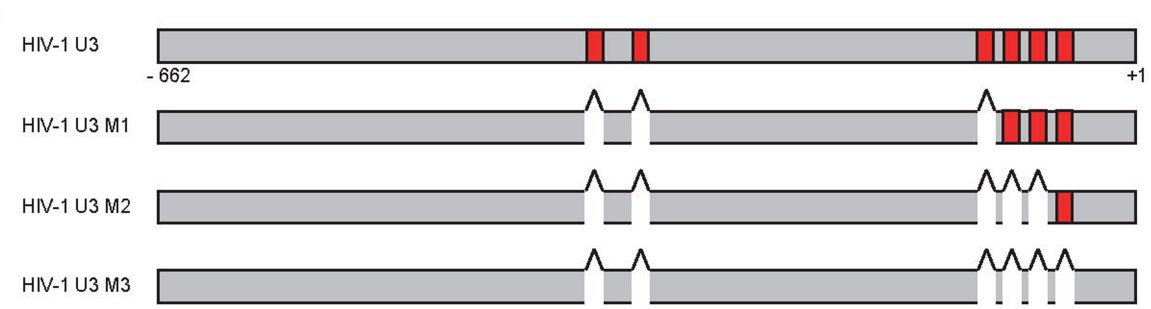
© 2018 Springer Nature Limited. All rights reserved.

a

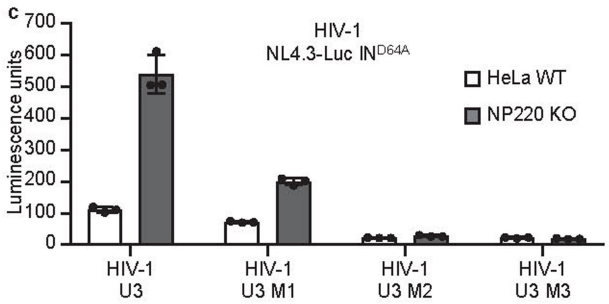
>>HIV-1 U3

AGGGCGGAAAGTCCAAATTGCTCGAGACCTAGAAAAACATGGAGCAATCACAAGTAGCAATACAGCAGCTAACAATGCTGCTTGTGCCTGGCTAGAAGCACAAGAGGAGGA
AGAGGTGGGTTTTCCAGTCACACCTCAGGTACCTTTAAGACCAATGACTTACAAGGCAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTAAT
TCACTCCCAAAGAAGACAAGATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTGGCAGAACTACACACCAGGGCCAGGGGTCAGATATCCACTGAC
CTTTGGATGGTGCTACAAGCTAGTACCAGTTGAGCCAGATAAGGTAGAAGAGGCCAATAAAGGAGAGAACACCAGCTTGTTACACCCTGTGAGCCTGCATGGAATGGATGA
CCCTGAGAGAGAAGTGTTAGAGTGGAGGTTTGACAGCCGCCTAGCATTTCATCACGTGGCCCGAGAGCTGCATCCGGAGTACTTCAAGAACTGCTGACATCGAGCTTGCTA
CAAGGGACTTTCCGCTGGGGACTTTCCAGGGAGGCGTGGCCTGGGCGGGACTGGGGAGTGGCGAGCCCTCAGATGCTGCATATAAGCAGCTGCTTTTTGCCTGTACT

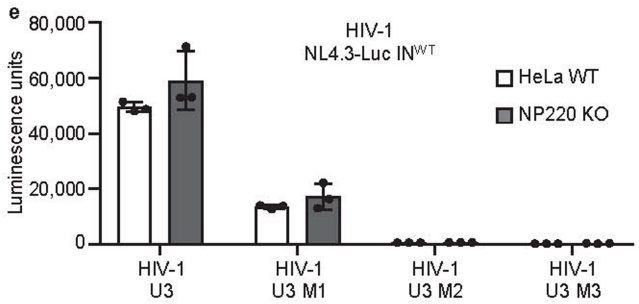
b



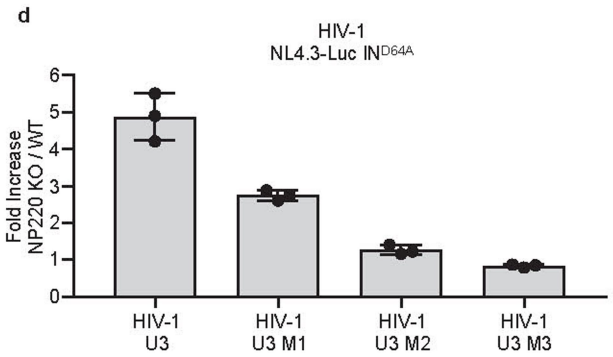
c



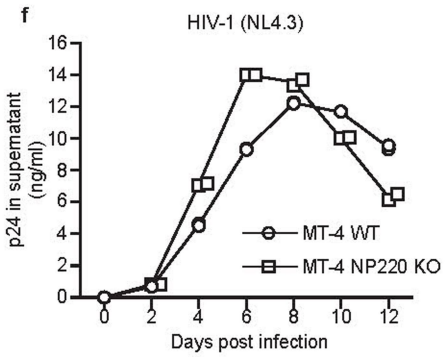
e



d

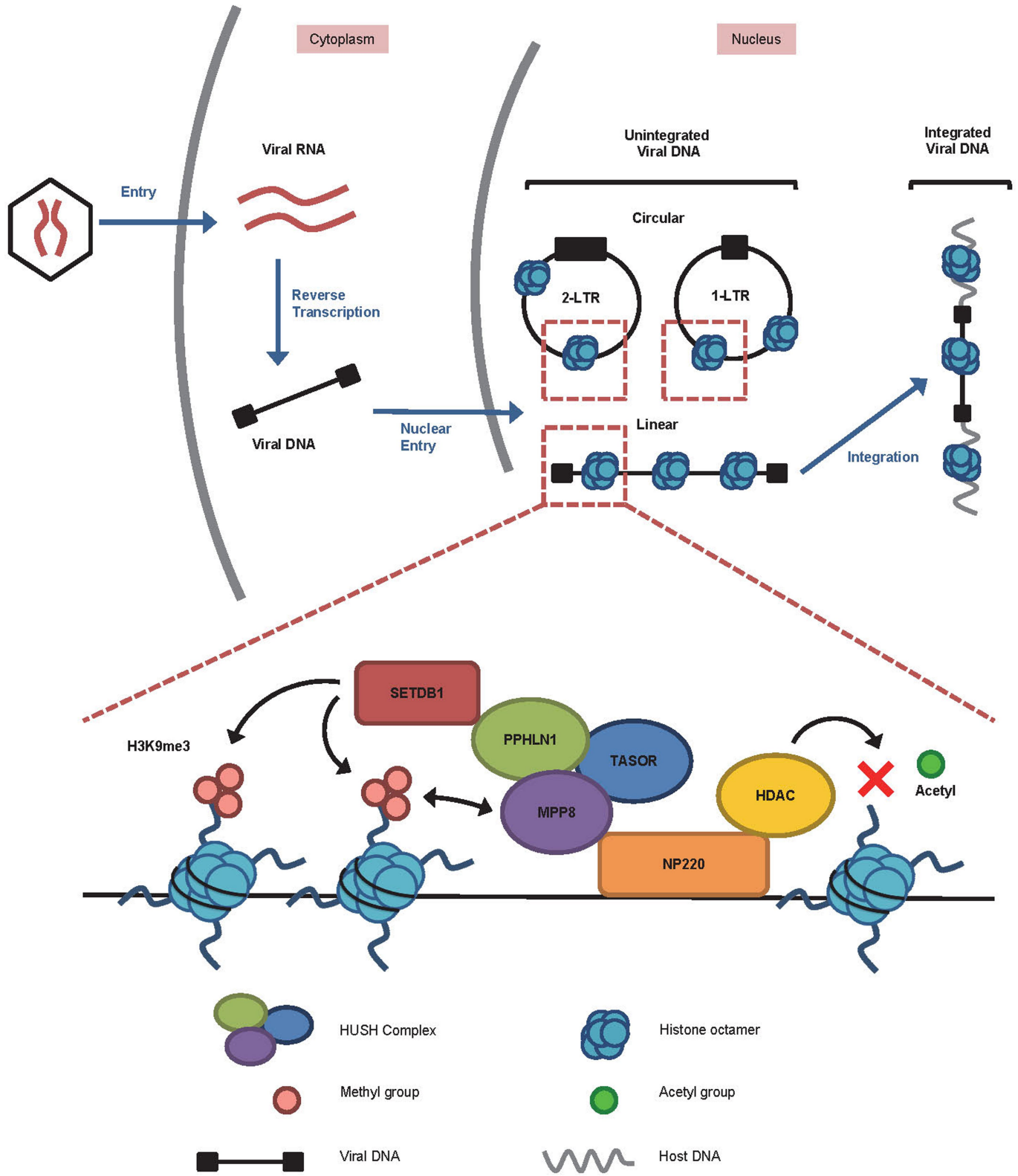


f



Extended Data Fig. 7 | The binding specificity of NP220 bound to unintegrated HIV-1 DNA. a, **b**, The sequences and locations of putative NP220-binding sites in HIV-1 U3 region. Putative NP220-binding sites are indicated in red. **c**, **d**, Parental or *NP220* knockout HeLa cells were infected with VSV-G pseudotyped, integrase-deficient (IN(D64A)) MLV-Luc viruses bearing indicated deletions in the U3 region. **c**, Luciferase activities were measured 40 h after infection. **d**, Fold increase (NP220 KO/NP220 WT) was calculated as the ratio of the luciferase activities of knockout cells compared to the luciferase activities of wild-type cells. Data are mean ± s.d.; $n = 3$ independent experiments. **e**, Parental or *NP220* knockout HeLa cells were infected with VSV-G pseudotyped, integrase-proficient (IN(WT)) HIV-Luc viruses bearing the indicated deletions in the U3 region. Luciferase activities were measured 40 h after infection. Data are mean ± s.d.; $n = 3$ independent experiments. **f**, Knockout of NP220 increases the rate of HIV-1 spreading. **g**, Parental and *NP220* knockout MT-4 cells were infected with HIV-1(NL4.3). Viral spreading was monitored by assay for p24 concentration in the culture medium. Data are mean from two technical ELISA replicates and are representative of two independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Schematic of the silencing of unintegrated retroviral DNAs. Retroviral infection results in the synthesis of a linear double-stranded DNA in the cytoplasm, which is delivered into the nucleus to give rise to two circular forms and the integrated provirus (top). The unintegrated nuclear DNAs are rapidly loaded with nucleosomal histones (blue). In the case of MLV, NP220 binds to the unintegrated viral DNA and is responsible for attracting histone deacetylases (HDACs), the HUSH complex (consisting of MPP8, TASOR and PPHLN1) and the histone methyltransferase SETDB1. HDACs remove the activation marks of histone acetylation and SETDB1 introduces repressive H3K9me3 marks. MPP8 binds H3K9me3 to strengthen the association with the viral chromatin.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0767-x

A chemical defence against phage infection

Sarah Kronheim[1], Martin Daniel-Ivad[1], Zhuang Duan[1], Sungwon Hwang[1], Andrew I. Wong[1], Ian Mantel[1], Justin R. Nodwell[1] & Karen L. Maxwell[1]*

The arms race between bacteria and the phages that infect them drives the continual evolution of diverse anti-phage defences. Previously described anti-phage systems have highly varied defence mechanisms[1–11]; however, all mechanisms rely on protein components to mediate defence. Here we report a chemical anti-phage defence system that is widespread in *Streptomyces*. We show that three naturally produced molecules that insert into DNA are able to block phage replication, whereas molecules that target DNA by other mechanisms do not. Because double-stranded DNA phages are the most numerous group in the biosphere and the production of secondary metabolites by bacteria is ubiquitous[12], this mechanism of anti-phage defence probably has a major evolutionary role in shaping bacterial communities.

The evolution of bacteria is strongly driven by their battle for survival against the viruses that infect them[13], which are known as phages. Phage infections are a major threat to bacteria and the resulting selective pressure has led to the emergence of diverse anti-phage defence mechanisms. These include cell-surface modifications that inhibit phage entry[1,2], abortive infection mechanisms that trigger cell death upon phage infection[3,4], restriction-modification[5] and CRISPR–Cas[6] systems that cleave invading phage genomes and many others[7–11]. The anti-phage defences described to date are highly varied in mechanism, but are all mediated by proteins or protein–RNA complexes. Here we report that bacterially produced small molecules, which are renowned for their ability to prevent microbial growth and serve as antibiotics, can also act as potent inhibitors of phage replication and represent a widespread anti-phage defence system.

Environmental bacteria produce a diverse repertoire of biologically active small molecules that are known as specialized or secondary metabolites[12]. Although these molecules are not essential for the survival of the organism[14], they confer a fitness advantage in competitive environments such as soil and marine sediments[15]. Many of these metabolites function as defence molecules, inhibiting the growth of other organisms that compete within the same niche[15,16]. Given the widespread role of secondary metabolites in protecting against cellular predators, we reasoned that some of them might defend against the other major bacterial enemy—phages. In addition, it was observed more than 50 years ago that *Streptomyces* species produce DNA-intercalating molecules that inhibit *Escherichia coli* phages[17,18], but the biological importance of these molecules was not recognized or investigated. Thus, we undertook a systematic investigation of the role of small molecules and secondary metabolites in the defence against phages.

Before initiating a search for naturally produced secondary metabolites that could serve as an anti-phage defence, we investigated how frequently small molecules inhibit phage replication. To this end, we used a high-throughput assay to screen a collection of 4,960 compounds—including natural products, FDA-approved drugs and known bioactive molecules—to identify compounds that protect *E. coli* from lysis by the well-characterized phage λ (Fig. 1a). In our assay, bacteria were first cultured for one hour in the presence of each compound. Phages were then added at a high multiplicity of infection and bacterial growth was monitored for five hours. In the absence of a compound, bacteria were completely lysed within two hours of phage addition. Phage-mediated lysis was also observed in the presence of most of the tested compounds (Fig. 1b). However, bacterial growth was observed in the presence of 11 compounds, indicating that these compounds were able to block phage replication and thus allow the bacteria to survive (Fig. 1c and Extended Data Table 1). Remarkably, 9 out of 11 compounds identified are DNA-intercalating agents and four of the compounds—daunorubicin, doxorubicin, epirubicin and idarubicin—are clinically used anti-cancer drugs. These four compounds fall into a class of molecules known as anthracyclines, which are secondary metabolites naturally produced by, and first isolated from, *Streptomyces*—a widespread genus of soil bacteria. The other intercalating molecules include a synthetic anthracycline (mitoxantrone), an alkaloid (ellipticine), a fluorochrome (propidium iodide) and acridine family compounds (acriflavine and ethacridine lactate). The remaining compounds are a di-benzimidazole (Ro 90-7501) and dequalinium chloride, a quaternary ammonium cation that increases bacterial cell permeability.

Streptomyces produce a huge number of diverse secondary metabolites, including more than two-thirds of clinically useful antibiotics[19]. Because phages threaten the survival of *Streptomyces* species in their natural soil environment, we reasoned that some of their secondary metabolites might exist for the purpose of anti-phage defence. This idea was bolstered by our identification of two *Streptomyces* secondary metabolites, daunorubicin and doxorubicin, in our screen for compounds that could block phage replication. To determine whether daunorubicin and doxorubicin might actually function 'in the wild' to protect *Streptomyces* from phage predation, we established a collection of twelve geographically diverse *Streptomyces* phages with distinct host ranges (Extended Data Fig. 1a). Electron microscopy revealed that they belong to the *Caudovirales* order, with long non-contractile tails and double-stranded DNA (dsDNA) genomes that are contained inside an icosahedral head (Extended Data Fig. 1b). We examined the ability of doxorubicin and daunorubicin to inhibit the replication of these phages in the well-characterized species, *Streptomyces coelicolor*. We determined an inhibitory concentration of 10 μM for daunorubicin and doxorubicin using phages ϕScoe25 and ϕScoe2, concentrations at which bacterial growth was not slowed (Extended Data Fig. 2a–c). In the absence of drug, the 12 phages were able to propagate robustly, producing about 10^8 phages per ml after overnight incubation. In the presence of 10 μM daunorubicin or doxorubicin, phage replication was inhibited 10^5-fold or more (Fig. 2a and Extended Data Fig. 2a, b). This demonstrates that both doxorubicin and daunorubicin, which are produced in the same biosynthetic pathway and differ by only a single hydroxyl group, can protect *Streptomyces* from phage predation.

To determine whether these compounds, when produced by their natural source (*Streptomyces peucetius*) could inhibit phage replication, cultures of this species were grown for 1, 2, 3 and 4 days in mannitol–soy broth to induce production. The characteristic red colour of doxorubicin and daunorubicin appeared at day 3 (Fig. 2b). Using mass spectrometry, we confirmed that doxorubicin and daunorubicin were present in the spent medium (Extended Data Fig. 3a, b). We found that doxorubicin predominated and was present at concentrations of 9 μM on day 3 and 18 μM on day 4, whereas the daunorubicin concentrations were 0.2 μM and 0.5 μM, respectively (Extended Data Fig. 2d, e).

[1]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada. *e-mail: karen.maxwell@utoronto.ca

© 2018 Springer Nature Limited. All rights reserved.

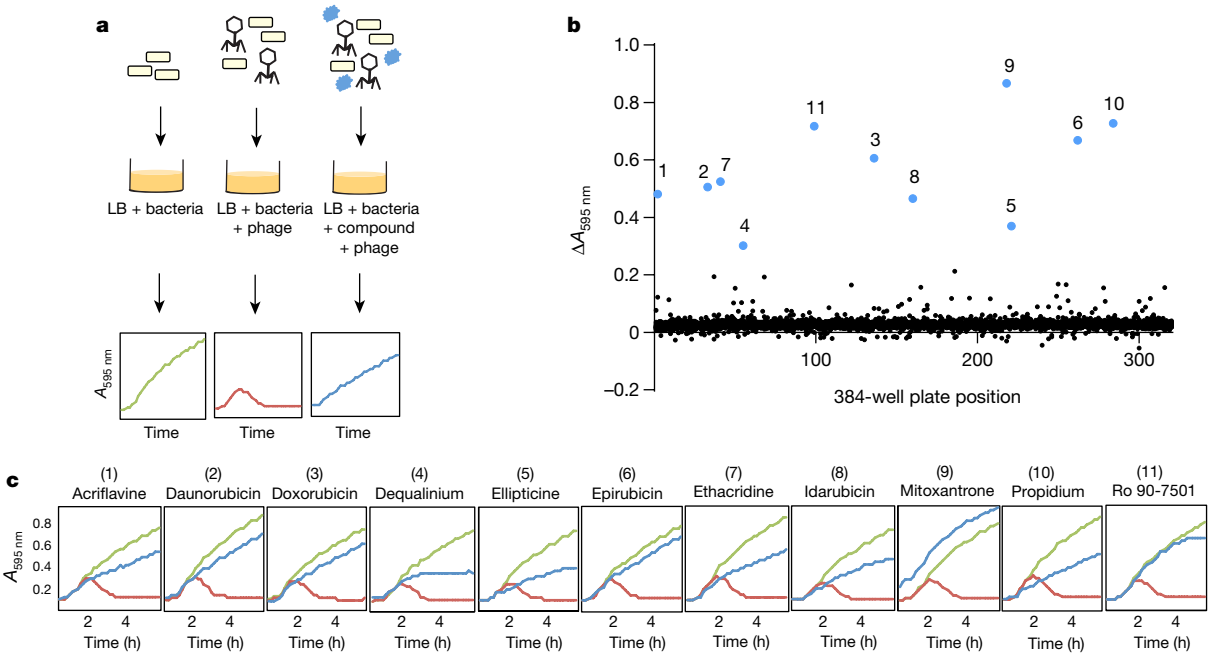


Fig. 1 | Small molecules inhibit phage replication. a, Anti-phage small-molecule screening protocol. Cultures of *E. coli* grown in the absence (green) or presence (red) of phage λ reveal phage-mediated lysis within 2 h of phage addition. Lysis can be inhibited by the addition of small molecules (blue). **b**, Compounds that result in cultures with a difference in absorbance at 595 nm ($\Delta A_{595\ nm}$) of >0.3 between time 0 and 5 h were considered to have anti-phage activity (11 compounds). The full drug library was screened once. **c**, Growth curves of phage λ infections in the presence of the 11 active anti-phage compounds, coloured as in **a**. $n = 3$ independent biological replicates. Numbers above growth curves indicate compounds shown in **b**.

Filter-sterilized spent medium from these cultures was mixed with fresh nutrients, *S. coelicolor* spores and phage ϕScoe2 and was incubated overnight. The following day, we determined the number of phages in each sample. We found that spent medium from days 3 and 4 robustly inhibited phage replication, whereas bacterial cell growth was not significantly affected (Fig. 2b and Extended Data Fig. 2c, f). The spent medium from days 1 and 2, which did not contain doxorubicin, did not inhibit phage replication. Similarly, spent medium from *S. peucetius* grown in nutrient broth, conditions which do not support doxorubicin production, also had no effect on phage replication. These data show that *S. peucetius* produces sufficient doxorubicin to prevent phage replication.

The production of secondary metabolites that inhibit phage replication would presumably deliver a considerable increase in fitness. Therefore, we expected this capability to be a conserved feature in diverse *Streptomyces* species. To investigate this possibility, we created small-molecule extracts from 48 *Streptomyces* strains from the Wright Actinomycete Collection (WAC)[20] and tested their ability to block the replication of two phages with distinct host ranges. Around 30% of the extracts (14 out of 48) inhibited the replication of one or both phages (Fig. 2c) by $\geq 10^4$-fold without preventing bacterial growth (Extended Data Fig. 4a). We set a high fold-inhibition cut-off to identify extracts with strong anti-phage activity and filter out those that might affect phage replication non-specifically—for example, through a global decrease in bacterial protein production. We subsequently tested these 14 extracts against our panel of twelve *Streptomyces* phages and found that some extracts (for example, WAC240) strongly inhibited the replication of all phages tested, whereas others (for example, WAC303 and WAC212) strongly inhibited only some phages (Fig. 2c). These data indicate that anti-phage secondary metabolites are commonly produced by *Streptomyces*. Furthermore, the variation in the phage inhibitory profiles, the varied geographical locations from which the *Streptomyces* strains were isolated and the metabolic diversity of these strains as determined by their analytical profile index[21] suggest that the extracts contain different anti-phage compounds (Extended Data Fig. 4b and Extended Data Table 2).

To directly address the question of whether distinct secondary metabolites were inhibiting phage replication, we analysed one extract with broad anti-phage activity (WAC240) and another that inhibited only some of the phages (WAC288). Using a bioactivity-guided fractionation approach, we isolated the active metabolites from each extract. Mass spectrometry of a pure active fraction of WAC288 showed that it contained cosmomycin D (Extended Data Fig. 5), an anthracyline that intercalates dsDNA as tightly as daunorubicin[22]. The same fractionation strategy revealed that the active compound in WAC240 was actinomycin D (Extended Data Fig. 6), another commonly used anti-tumour agent. This molecule also has DNA intercalating activity, but is not an anthracycline[23]. We confirmed their activities in our phage replication assay in *S. coelicolor* using commercial compounds (Fig. 2a and Extended Data Fig. 2g). To determine whether any of the twelve remaining active extracts contained daunorubicin, cosmomycin D, actinomycin or related compounds, we mapped their small molecule contents using liquid chromatography–tandem mass spectrometry and searched for mass fragments that correspond to these compounds. We found that the extract from WAC218 contained cosmomycin B, and WAC205 and WAC251 extracts contained fragments that match the mass of the core anthracycline moiety of daunorubicin. These data highlight the frequent occurrence of DNA-intercalating compounds in phage-inhibiting extracts and suggest that DNA intercalation may represent a particularly effective means to block phage replication. Consistent with this finding, our original screen in *E. coli* identified two other distinct DNA-intercalating molecules—propidium iodide and acriflavine—that inhibit phage replication. We found that these two DNA-intercalating compounds also inhibited *Streptomyces* phages (Fig. 2a). To determine whether DNA intercalation is the key feature of these anti-phage molecules, we tested other small molecules that target DNA. Synthetic antibiotics that induce double-stranded breaks through their interaction with DNA gyrase—such as ciprofloxacin and nalidixic acid—did not inhibit phage replication (Fig. 2a). The *Streptomyces*-derived anti-cancer drugs bleomycin and mitomycin C, which cleave and crosslink DNA, respectively, also lacked this activity (Fig. 2a). Therefore, DNA

© 2018 Springer Nature Limited. All rights reserved.

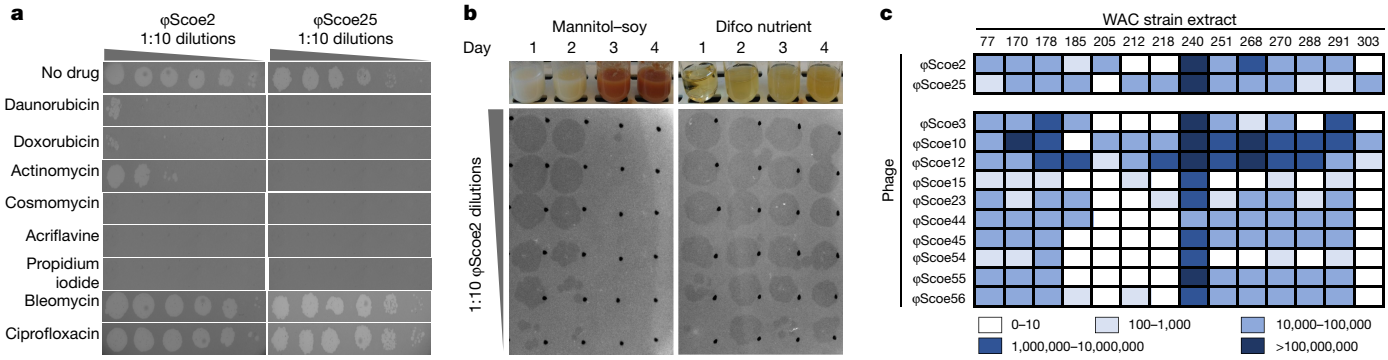


Fig. 2 | Secondary metabolites generated by *Streptomyces* inhibit phage replication. a, Propagation assays of phages ϕScoe2 and ϕScoe25 reveal that molecules that insert into DNA strongly inhibit phage replication, whereas other DNA-interacting molecules do not. $n = 3$ biological replicates. **b**, Doxorubicin production (red colour) is induced by growth in mannitol–soy broth (top). Phage replication assays in the presence of spent medium shows the inhibitory activity of doxorubicin (bottom). Growth in Difco nutrient broth did not induce doxorubicin production or inhibit phage propagation. $n = 3$ biologically independent experiments. **c**, Fourteen *Streptomyces* extracts inhibit phage replication up to $>10^8$-fold. $n = 5$ assays on three extract preparations.

intercalation appears to be the crucial property for anti-phage activity, rather than a general ability to damage DNA.

The remarkable ability of DNA-intercalating agents to block different phages in two widely divergent bacterial species led us to investigate the inhibitory mechanism of this class of molecules. Owing to technical limitations of the *Streptomyces* phage assay system, we assessed the effects of daunorubicin on the replication of *E. coli* phage λ (Fig. 3a and Extended Data Fig. 7a). The life cycle of a phage initiates with injection of the genome into the host cell. We used a potassium efflux assay[1,24,25] to test the ability of phage λ to inject its DNA into the host cell in the presence and absence of daunorubicin. Potassium release—correlating with successful delivery of the phage genome into the cytoplasm of the bacterial cell—was observed in both cases (Fig. 3b), indicating that the drug does not prevent the phage genome from entering the cell. We found that induction of a λ prophage (a phage genome integrated into the bacterial chromosome) in the presence of daunorubicin caused no reduction in the release of phage particles (Fig. 3c and Extended Data Fig. 7b), showing that the drug does not inhibit phage replication, protein synthesis or assembly of the viral particle. Pre-incubation of phage λ with daunorubicin did not inactivate the phage particle (Fig. 3c and Extended Data Fig. 7b). Finally, we examined the ability of phage λ to integrate into the bacterial chromosome as a lysogen. In this assay, a single infection event is examined and no replication is required. Using a high multiplicity of infection so that 100% of the bacterial cells were lysogenized in the absence of daunorubicin, we found a 70% reduction in phage λ lysogen formation in the presence of daunorubicin (Fig. 3d). Taken together, these experiments indicate that daunorubicin blocks a very early step that occurs after DNA injection, but before DNA replication.

At this time, we can only speculate as to the mechanism by which sub-inhibitory concentrations of daunorubicin and other intercalating compounds selectively inhibit phage replication without affecting bacterial growth. Upon infection, the phage genome is distinct from the bacterial genome in that it is linear, non-supercoiled and unprotected by DNA-binding proteins[26]. At this stage, the phage DNA may be uniquely sensitive to the effect of low levels of DNA-intercalating agents. These agents may prevent DNA circularization or interaction with the proteins that are required for replication and transcription. Consistent with previous studies[17,18], we observed that daunorubicin was able to block the replication of a variety of *E. coli* dsDNA phages. We found that all types of dsDNA phages, including those belonging to the siphophage, myophage and podophage families were inhibited by daunorubicin, whereas the single-stranded DNA (ssDNA) phage M13 was not (Table 1). We tested two *Pseudomonas aeruginosa* phages and found that these phages were inhibited as well (Table 1). Because

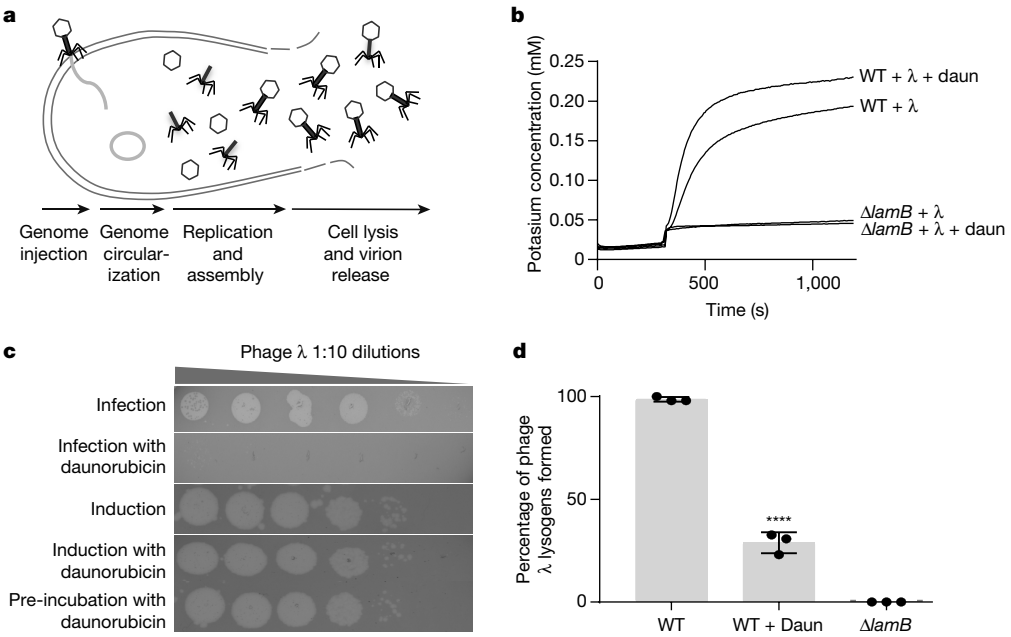


Fig. 3 | Daunorubicin acts at an early stage of phage λ infection. a, The life cycle of a phage presents a number of steps that could be inhibited by secondary metabolites. **b**, Potassium efflux assays with wild-type (WT) cells or cells that lack the λ receptor (Δ*lamB*) reveal that daunorubicin does not block phage genome entry. $n = 3$ independent biological replicates. **c**, Induction of a λ lysogen in the presence of daunorubicin or pre-treatment of the phage does not inhibit phage plaque formation (see also Extended Data Fig. 6b). $n = 3$ biological replicates. **d**, Daunorubicin decreases phage λ lysogen formation by approximately 70%. $n = 3$ independent biological replicates. $^{****}P < 0.0001$; analysis of variance (ANOVA).

© 2018 Springer Nature Limited. All rights reserved.

Table 1 | Effects of daunorubicin on propagation of *E. coli* and *P. aeruginosa* phages

Phage	Fold reduction in titre	Host	Phage type	Genome	Incoming genome form[26]
λ	10^{10}	*E. coli*	Siphophage	dsDNA; 48.5 kB	Linear with cohesive ends
T5	10^{7}	*E. coli*	Siphophage	dsDNA; 121.7 kB	Linear with terminal redundancy
T6	10^{3}	*E. coli*	Myophage	dsDNA; 168.9 kB	Linear with terminal repeats
T7	10^{4}	*E. coli*	Podophage	dsDNA; 39.9 kB	Linear with terminal redundancy
M13	0	*E. coli*	Filamentous	ssDNA; 6.4 knt	(+) strand
JBD26	100	*P. aeruginosa*	Siphophage	dsDNA: 37.8 kB	Host DNA at termini
JBD30	10	*P. aeruginosa*	Siphophage	dsDNA: 36.9 kB	Host DNA at termini

dsDNA phages are by far the most abundant type of phage in the environment[27], this mechanism of protection would be highly selected through evolution.

Bacterial anti-phage defence systems, known to be highly diverse and widespread in nature, were previously shown to rely on protein components to mediate resistance. Here we demonstrate that *Streptomyces* species commonly produce secondary metabolites that provide them with a chemical anti-phage defence system. In contrast to other bacterial anti-phage systems (which are usually specific to a narrow range of phages), small molecules such as DNA-intercalating compounds could provide broad protection against all dsDNA phages, by targeting an essential and universal step early in the life cycle of a phage. Because many secondary metabolites are known to be diffusible molecules, in the context of a bacterial community it could provide an innate defence system to protect the community as a whole.

In addition to the discovery of an anti-phage defence mechanism, this work also illuminates an alternative role for bacterially produced secondary metabolites. Because these metabolites are produced by most bacterial species, this anti-phage defence is likely to be widespread in nature. Furthermore, because the *Streptomyces* genus is predicted to be capable of producing on the order of 100,000 antimicrobial compounds[28], most of which are as-yet uncharacterized, we expect that other classes of anti-phage molecules also exist. Identification of these compounds will reveal new groups of bioactive compounds that, similar to the anthracyclines used in cancer treatment, may have broad applicability for therapeutic purposes. Therefore, further investigations to identify and characterize these metabolites and their mechanisms of action will uncover unique chemical matter, expand our knowledge of bacterial anti-phage defence systems and improve our understanding of the evolutionary forces that shape bacterial communities.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0767-x.

Received: 25 May 2018; Accepted: 2 November 2018;
Published online 5 December 2018.

1. Cumby, N., Edwards, A. M., Davidson, A. R. & Maxwell, K. L. The bacteriophage HK97 gp15 moron element encodes a novel superinfection exclusion protein. *J. Bacteriol.* **194**, 5012–5019 (2012).
2. Scholl, D., Adhya, S. & Merril, C. *Escherichia coli* K1's capsule is a barrier to bacteriophage T7. *Appl. Environ. Microbiol.* **71**, 4872–4874 (2005).
3. Molineux, I. J. Host–parasite interactions: recent developments in the genetics of abortive phage infections. *New Biol.* **3**, 230–236 (1991).
4. Chopin, M. C., Chopin, A. & Bidnenko, E. Phage abortive infection in lactococci: variations on a theme. *Curr. Opin. Microbiol.* **8**, 473–479 (2005).
5. Tock, M. R. & Dryden, D. T. The biology of restriction and anti-restriction. *Curr. Opin. Microbiol.* **8**, 466–472 (2005).
6. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
7. Goldfarb, T. et al. BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183 (2015).
8. Ofir, G. et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90–98 (2018).
9. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
10. Stern, A. & Sorek, R. The phage–host arms race: shaping the evolution of microbes. *BioEssays* **33**, 43–51 (2011).
11. Seed, K. D. Battling phages: how bacteria defend against viral attack. *PLoS Pathog.* **11**, e1004847 (2015).
12. Davies, J. Specialized microbial metabolites: functions and origins. *J. Antibiot.* **66**, 361–364 (2013).
13. Koskella, B. & Brockhurst, M. A. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* **38**, 916–931 (2014).
14. Komatsu, M., Uchiyama, T., Omura, S., Cane, D. E. & Ikeda, H. Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism. *Proc. Natl Acad. Sci. USA* **107**, 2646–2651 (2010).
15. Vetsigian, K., Jajoo, R. & Kishony, R. Structure and evolution of *Streptomyces* interaction networks in soil and in silico. *PLoS Biol.* **9**, e1001184 (2011).
16. Barka, E. A. et al. Taxonomy, physiology, and natural products of *Actinobacteria*. *Microbiol. Mol. Biol. Rev.* **80**, 1–43 (2015).
17. Parisi, B. & Soller, A. Studies on the antiphage activity of daunomycin. *Giorn. Microbiol.* **12**, 183–194 (1964).
18. Morita, J., Tanaka, A., Komano, T. & Oki, T. Inactivation of phage ϕX174 by anthracycline antibiotics, alacinomycin A, doxorubicin and daunorubicin. *Agric. Biol. Chem.* **43**, 2629–2631 (1979).
19. Lucas, X. et al. StreptomeDB: a resource for natural compounds isolated from *Streptomyces* species. *Nucleic Acids Res.* **41**, D1130–D1136 (2013).
20. Thaker, M. N. et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **31**, 922–927 (2013).
21. Smith, P. B., Tomfohrde, K. M., Rhoden, D. L. & Balows, A. API system: a multitube micromethod for identification of Enterobacteriaceae. *Appl. Microbiol.* **24**, 449–452 (1972).
22. Furlan, R. L. et al. DNA-binding properties of cosmomycin D, an anthracycline with two trisaccharide chains. *J. Antibiot.* **57**, 647–654 (2004).
23. Waksman, S. A., Geiger, W. B. & Reynolds, D. M. Strain specificity and production of antibiotic substances. VII. Production of actinomycin by different Actinomycetes. *Proc. Natl Acad. Sci. USA* **32**, 117–120 (1946).
24. Boulanger, P. & Letellier, L. Ion channels are likely to be involved in the two steps of phage T5 DNA penetration into *Escherichia coli* cells. *J. Biol. Chem.* **267**, 3168–3172 (1992).
25. Boulanger, P. & Letellier, L. Characterization of ion channels involved in the penetration of phage T4 DNA into *Escherichia coli* cells. *J. Biol. Chem.* **263**, 9767–9775 (1988).
26. Casjens, S. R. & Gilcrease, E. B. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol. Biol.* **502**, 91–111 (2009).
27. Ackermann, H. W. 5500 phages examined in the electron microscope. *Arch. Virol.* **152**, 227–243 (2007).
28. Watve, M. G., Tickoo, R., Jog, M. M. & Bhole, B. D. How many antibiotics are produced by the genus *Streptomyces*? *Arch. Microbiol.* **176**, 386–390 (2001).

Acknowledgements We thank G. Wright for providing the WAC strains and D. Bona for technical assistance. We thank A. Davidson for many useful discussions and comments on the manuscript, M. Meneghini and W. Navarre for reading the manuscript and A. Edwards for his long-standing support. This work was supported by operating grants from the Natural Sciences and Engineering Research Council (RGPIN-2017-06898) and Canadian Institutes of Health Research (CIHR; MOP-136845) to K.L.M. and CIHR (MOP-57684) to J.R.N. M.D.-I. was supported by a Studentship from the Canadian Cystic Fibrosis Foundation.

Reviewer information *Nature* thanks M. Clokie and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.K. performed and analysed most of the biological experiments, M.D.-I. performed mass spectrometry and analytical profile index 20E analyses, Z.D. performed the high-throughput *E. coli* drug screen, and S.H., A.I.W., and I.M. assisted with phage isolation and plaque experiments. J.R.N. and K.L.M. supervised the study.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0767-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0767-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.L.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

High-throughput screen for anti-phage compounds in *E. coli*. A collection of approved, bioactive, natural product compounds purchased from the SPARC Biocentre (http://lab.research.sickkids.ca/sparc-drug-discovery/) was screened for the ability to inhibit the replication of *E. coli* phage λ. This collection includes the LOPAC^{1280} (1,280 compounds; Sigma-Aldrich), Prestwick Chemical Library (1,280 compounds; Prestwick Chemical) and The Spectrum Collection (2,400 compounds; MicroSource Discovery Systems) libraries. To screen the compounds for anti-phage activity, *E. coli* strain BW25113 (Coli Genetic Stock Center) was grown overnight with shaking in lysogeny broth (LB) medium at 37 °C. A 2% subculture was prepared in LB medium and 50 μl was transferred to the wells of a 384-well flat-bottom plate. Each well contained one of the LOPAC, Prestwick or MicroSource compounds at a final concentration of 40 μM. Plates were incubated with orbital shaking at an amplitude of 3 mm in a Tecan Infinite 200 plate reader at 37 °C, with $A_{595\,nm}$ readings were taken every 15 min. At 1 h, plates were taken from the plate reader and phage λcI$_{857}$ was added at a multiplicity of infection of 10. The plates were returned to the plate reader and the $A_{595\,nm}$ was recorded every 15 min for an additional 5 h. The $A_{595\,nm}$ of each well in the 384-well plate at time 0 was subtracted from the $A_{595\,nm}$ after 5 h of incubation, and the resulting numbers were plotted as a function of well number. A positive result was defined as a change in $A_{595\,nm}$ greater than 0.3, which signified a compound that prevented λcI$_{857}$-mediated bacterial lysis and allowed bacterial cell growth. We chose this cut-off to identify compounds that had robust anti-phage activity and were not acting indirectly, for example, through a global decrease in bacterial protein production.

Preparation of high-titre *Streptomyces* spore stocks. In brief, 10 μl of frozen spore stocks were inoculated on MYM (4 g l^{-1} maltose, 4 g l^{-1} yeast extract and 10 g l^{-1} malt extract; supplemented with 2 ml trace elements following autoclaving) agar plates and incubated for five days at 30 °C to allow growth and development of *Streptomyces* spores. Once spores developed, 10 ml of 0.85% NaCl was added to the plate and spores were removed by gentle scraping. The saline and spores were collected, filtered through a sterile cotton ball, and centrifuged at 3,000 r.p.m. for 10 min. Pellets were resuspended in saline with 15% glycerol and stored at −80 °C.

Isolation of *Streptomyces* phages. A small scoop of soil from various locations was added to 20 ml of Difco nutrient broth (DNB) medium supplemented with 0.5% glucose and 4 mM $CaCl_2$. Spores of *Streptomyces avermitilis* SUKA22, *S. coelicolor* M145 or WAC212 were added to a final concentration of approximately 10^7 spores per ml and the culture was incubated overnight at 30 °C with shaking. The following morning, the cultures were pelleted by centrifugation to remove cells and bacterial debris, and the supernatant was filter-sterilized. The filtrate was diluted and plated onto DNB agar overlaid with the appropriate spores suspended in 0.5% (w/v) top agar. Plates were incubated at 30 °C overnight, then single plaques were selected and purified through three rounds of plaque purification.

Determination of the host range of isolated *Streptomyces* phages. Dilutions of each phage were spotted on DNB plates overlaid with 0.5% (w/v) top agar containing spores of the relevant bacteria. The species assessed were: *S. avermitilis* SUKA22, *S. coelicolor* M145, *Streptomyces scabiei*, *Streptomyces venezuelae*, *Streptomyces viridochromogenes*, WAC165, WAC183, WAC212, WAC240 and WAC288. The host ranges were confirmed by two repetitions of spotting single dilutions of each phage on the bacterial strains.

Analysis of *Streptomyces* phages via transmission electron microscopy. High-titre phage lysates were prepared by adding approximately 10^7 spore forming units of *S. coelicolor* M145 spore stock and approximately 10^5 plaque forming units of each phage to 100 ml of Difco nutrient broth. The cultures were grown overnight at 30 °C, the cells were pelleted by centrifugation and the phage lysate was filtered through a 0.22-μm filter. The phage lysate was concentrated to approximately 1 ml using a GE Healthcare Vivaspin 20 concentrator with 100 kDa MWCO. Transmission electron microscopy (TEM) grids (Electron Microscopy Sciences CF400-CU) were prepared by drop-coating grids with each phage, washing with water and staining with 2% uranyl acetate. Phages were imaged using the Talos L120C transmission electron microscope in the Microscopy Imaging Laboratory at the University of Toronto.

Phenotypic characterization of *Streptomyces* isolates. To confirm that the WAC strain isolates belonged to the *Streptomyces* genus, cells were grown in MYM liquid culture for three days at 30 °C. Cells were collected by centrifugation and genomic DNA was extracted using the DNeasy kit (Qiagen). 16S rRNA sequences were amplified with primers (27f 5′-GAGTTTGATCCTGGCTCA-3′, 1492r 5′-TACGGCTACCTTGTTACGACTT-3′) designed to capture the full sequence. The phenotypic characterizations of each strain were performed using the analytical profile index 20E test by bioMerieux according to their protocol. These provide a complete biochemical profile of each strain (Extended Data Table 2).

Preparation of butanol cell-free specialized metabolite extracts. Extracts of the strains from the WAC (http://www.thewrightlab.com/wright-actinomycete-collection/) were prepared by culturing strains on solid MYM medium for five days at 30 °C. The agar was removed from the plates, cut into small cubes and immersed in butanol overnight. This extract was filtered through Whatman paper and the filtrate was dried in vacuo. Small-scale extracts for preliminary analysis and confirmation of activity were made from one 10-cm Petri dish, each resuspended in DMSO to a final concentration of 80 mg ml^{-1}. Large-scale extracts for purification of active compounds were made from batches of 20 MYM plates and resuspended in 1:1 acetonitrile:water.

Analysis of anti-phage activity of *Streptomyces* extracts and purified compounds. One small glass bead and 200 μl of DNB medium supplemented with 0.5% (w/v) glucose and 4 mM $CaCl_2$ were added to the wells of a 96-well plate. Each well was inoculated with 10^5 spores of *S. coelicolor* M145, phages at a multiplicity of infection of 0.1 and the indicated compound or extract. Final concentrations were 1 mg ml^{-1} for WAC extracts, 0.2 mg ml^{-1} purified cosmomycin D, 10 μM daunorubicin, 10 μM doxorubicin, 40 μM actinomycin, 40 μM acriflavine, 40 μM propidium iodide, 40 μM bleomycin, 10 μM mitomycin C, 40 μM nalidixic acid and 5 μM ciprofloxacin. To mitigate effects generated by the edge of the plate, sterile water was added to each empty well bordering a well that contained the propagation mixture to limit evaporation. Plates were incubated overnight with shaking at 30 °C. The number of phage particles that were present after overnight propagation was determined by collecting each culture supernatant and spotting serial dilutions onto a lawn of *S. coelicolor* M145 on DNB plates. Plaques were enumerated following overnight incubation at 30 °C. Three biological replicates were performed for the purified drug assays. Four biological replicates were performed for the WAC activity screening. An extract was deemed active if it showed activity in at least three out of four replicates.

Monitoring anti-phage activity of *S. peucetius* supernatants. *S. peucetius* was cultured in DNB and maltose–soy media and incubated with shaking at 220 r.p.m. with glass beads at 30 °C for up to four days. To determine whether daunorubicin or doxorubicin was present in the cultures, supernatants were clarified through a 0.2-μm nylon filter (Pall) and injected onto a BEH C18 2.1 mm × 100 mm (1.7 μm) UPLC column (Waters). Separation was accomplished with a linear gradient from 5% to 95% water:acetonitrile over 25 min at 40 °C. Analytes were detected by a G2S-qTof high-resolution mass spectrometer (Waters). To ascertain whether anti-phage activity was present, 500 μl of each filtered culture supernatant was mixed with 10× DNB medium supplemented with 5% glucose (w/v), 40 mM $CaCl_2$, 10^5 spores of *S. coelicolor* M145 and 10^4 phage particles. The culture was incubated overnight at 30 °C with shaking. The resulting culture supernatant was collected and serial dilutions were spotted onto a lawn of *S. coelicolor* M145. Plates were incubated at 30 °C overnight and phages were enumerated by counting plaques. It should be noted that phage-replication assays could not be performed directly in *S. peucetius* owing to its very slow growth rate and delayed production of daunorubicin.

Monitoring production of daunorubicin and doxorubicin in *S. peutecius*. Culture supernatants of *S. peucetius* grown in either DNB or mannitol–soy media were clarified with a 0.2-μm nylon filter (Pall) and injected onto a BEH C18 2.1 mm × 100 mm (1.7 μm) UPLC column (Waters). Separation was accomplished with a linear gradient from 5% to 95% water:acetonitrile over 25 min at 40 °C and analytes were detected by a G2S-qTof high-resolution mass spectrometer (Waters). Daunorubicin and doxorubicin were quantified by generating a standard curve with known concentrations of commercially available standards of each. Ion counts were quantified from the proton adducts of daunorubicin and doxorubicin with a mass error window of ±0.2 Da repeated in triplicate. Quantification of analytes in the supernatants was conducted in the same manner as above and ion counts were compared to the standard curve to calculate the concentration (Extended Data Fig. 2).

Determining bactericidal activity of compounds and extracts. *S. coelicolor* M145 was grown overnight at 30 °C in 500 μl of DNB medium in the presence of the indicated compound or extract. Bacterial growth was assessed by plating serial dilutions of the cell culture following vigorous vortexing to separate clumps of cells on DNB plates and incubating the plates for 48 h at 30 °C. Colonies were counted at the lowest dilution at which individual colonies were present. This was performed for each active WAC extract at a final concentration of 1 mg ml^{-1}. The working concentrations of actinomycin, cosmomycin D, acriflavine, propidium iodide, bleomycin, ciprofloxacin, nalidixic acid and mitomycin C were determined in *S. coelicolor* M145 by growing cells overnight in liquid MYM with each compound (except cosmomycin D) at concentrations of 40, 20, 10, 5 and 2.5 μM. Serial dilutions of the overnight growth were spotted on plates, incubated for 30 h to allow colonies to grow and colony-forming units were determined. Concentrations of cosmomycin D purified from WAC288 ranging from 0.8 mg ml^{-1} to 0.05 mg ml^{-1} were tested in the same manner. Three biological replicates were grown and analysed for each compound and concentration as well as the control condition

© 2018 Springer Nature Limited. All rights reserved.

that did not contain any added compounds. The highest concentration of each compound that did not inhibit the growth of *S. coelicolor* M145 was used for the remainder of the study. The concentrations that were used for the phage propagations were 40 μM actinomycin, 40 μM acriflavine, 40 μM propidium iodide, 40 μM bleomycin, 40 μM nalidixic acid, 10 μM mitomycin C, 5 μM ciprofloxacin and 0.2 mg ml^{-1} cosmomycin D.

Purification and characterization of cosmomycin D. Cosmomycin D was isolated by first inoculating a seed culture of 10 ml MYM medium with WAC288 spores at 30 °C for four days. This culture was used to inoculate five 50-ml cultures of R5M (10 g l^{-1} glucose, 5 g l^{-1} yeast extract, 0.1 g l^{-1} casamino acids, 10 g l^{-1} magnesium chloride hexahydrate; after autoclaving supplemented with 2.5 ml 0.5% (w/v) potassium phosphate, 1 ml 5 M calcium chloride and 2 ml trace elements) in 250-ml baffled flasks. These production cultures were grown for an additional six days at 30 °C, with shaking at 220 r.p.m. Culture supernatants were clarified by centrifugation (3,000*g* for 10 min), mixed with 20 g l^{-1} Diaion HPSS20 resin and incubated overnight. The extract was packed in a flash cartridge and separated using a linear water:methanol gradient over 1 h at 5 ml min^{-1}. Cosmomycin-D-containing fractions were collected and dried in vacuo. Samples were further purified using a Luna C18 (4.6 mm × 250 mm, 5-μm particle) high-performance liquid chromatography (HPLC) column (Phenomenex) using a water:acetonitrile gradient supplemented with 0.1% formic acid in both mobile phases flowing at 1 ml min^{-1} at 35 °C. The purification gradient follows a hold of 10% acetonitrile for 5 min, holding 20% acetonitrile for 7 min, holding 30% acetonitrile for 8 min and a linear gradient to 95% for 15 min, monitoring at 494 nm (reflective of the anthracycline moiety). The cosmomycin D structure was confirmed by high-resolution tandem mass spectrometry.

Purification and characterization of actinomycin D and C. A lawn of WAC240 was plated on solid MYM medium and grown for seven days at 30 °C. Following incubation, the plates were sectioned into cubes and macerated overnight in butanol. The extract was filtered and dried in vacuo. The crude extract was resuspended in 1:1 methanol and water and separated on a C18 flash cartridge (Agela Technologies) using a sequential 25% step gradient flowing at 15 ml min^{-1}. The active fractions were dried and further purified by HPLC on a Luna C18 (10 mm × 250 mm, 5-μm particle) column (Phenomenex) using a 5 ml min^{-1} linear gradient of water and acetonitrile from 20% to 95% acetonitrile over 20 min, monitoring at 444 nm. The resulting active fraction was subjected to subsequent isolation by an analytical Luna C18 (4.6 mm × 250 mm, 5-μm particle, Phenomenex) using a linear gradient of 50% to 95% acetonitrile over 9 min. The presence of actinomycin D and C was confirmed by high-resolution tandem mass spectrometry.

Analysis of anti-phage activity of daunorubicin on *E. coli* phage λ. Cultures of *E. coli* BW25113 were grown in LB medium at 37 °C overnight. Subsequently, 1% subcultures in LB supplemented with daunorubicin (40 μM–2.5 μM) were grown at 37 °C for 1 h and phage λ was added at a multiplicity of infection of approximately 10. Cultures were grown for 6 h at 37 °C and the number of phage particles present after a 6-h incubation was determined by collecting each culture supernatant and spotting serial dilutions onto a lawn of *E. coli* on LB plates. Three biological replicates were tested for each concentration of daunorubicin as well as the control, which contained no daunorubicin.

Potassium efflux assays. Cultures of *E. coli* BW25113 and BW25113 *lamB* mutant (Δ*lamB*) cells, which lack the outer membrane receptor for phage λ, were grown in LB medium at 37 °C overnight. Subsequently, 1% subcultures in LB supplemented with 15 μM daunorubicin were grown at 37 °C to $A_{600\,nm} = 0.5$. Then, 3-ml aliquots of cells were collected via centrifugation, washed with suspension medium 3 (SM3; 10 mM Tris pH 7.5, 10 mM NaCl, 4 mM $MgSO_4$) and resuspended in 3 ml of SM3. Daunorubicin was added to SM3 to a final concentration of 15 μM. Cells were equilibrated at 37 °C for 5 min before the addition of 100 μl of phage λ (10^{12} plaque forming units per ml). Potassium levels in the bulk medium were monitored for 15 min with an Orion ionplus potassium electrode (Thermo Scientific).

Phage lysogen-formation assay. Cultures of *E. coli* BW25113 and BW25113 Δ*lamB* cells were grown in LB medium supplemented with 0.2% (w/v) maltose and 10 mM magnesium sulfate at 37 °C overnight with shaking. Cells were diluted to 1% (v/v) in the same medium in the presence of 15 μM daunorubicin and grown at 37 °C to $A_{600\,nm} = 0.5$. Phage λ with a temperature-sensitive repressor protein (λcI_{857}) was added to the cells at a multiplicity of infection of 10 and incubated at 30 °C for 15 min to allow phage infection. This mixture was diluted 1:1 with LB medium and recovered for 1 h at 30 °C. Dilutions of the cell culture were plated on LB plates to individual colonies and incubated at 30 °C overnight. On the following day, 52 colonies from each sample were plated in replicate. One plate was incubated overnight at 30 °C and the other at 42 °C. Lysogens were counted as colonies that could grow at 30 °C but not at 42 °C, owing to induction of the temperature-sensitive λ lysogen. This experiment was performed in triplicate.

Pre-incubation of phage λ with daunorubicin. Aliquots of phage λ were incubated at 37 °C for 2 h with 15 μM daunorubicin. The phages were diluted with LB medium and serial dilutions were spotted onto a lawn of *E. coli* BW25113 cells suspended in molten top agar overlaid onto LB plates. The plates were incubated at 37 °C overnight and active phage particles were enumerated by counting plaques.

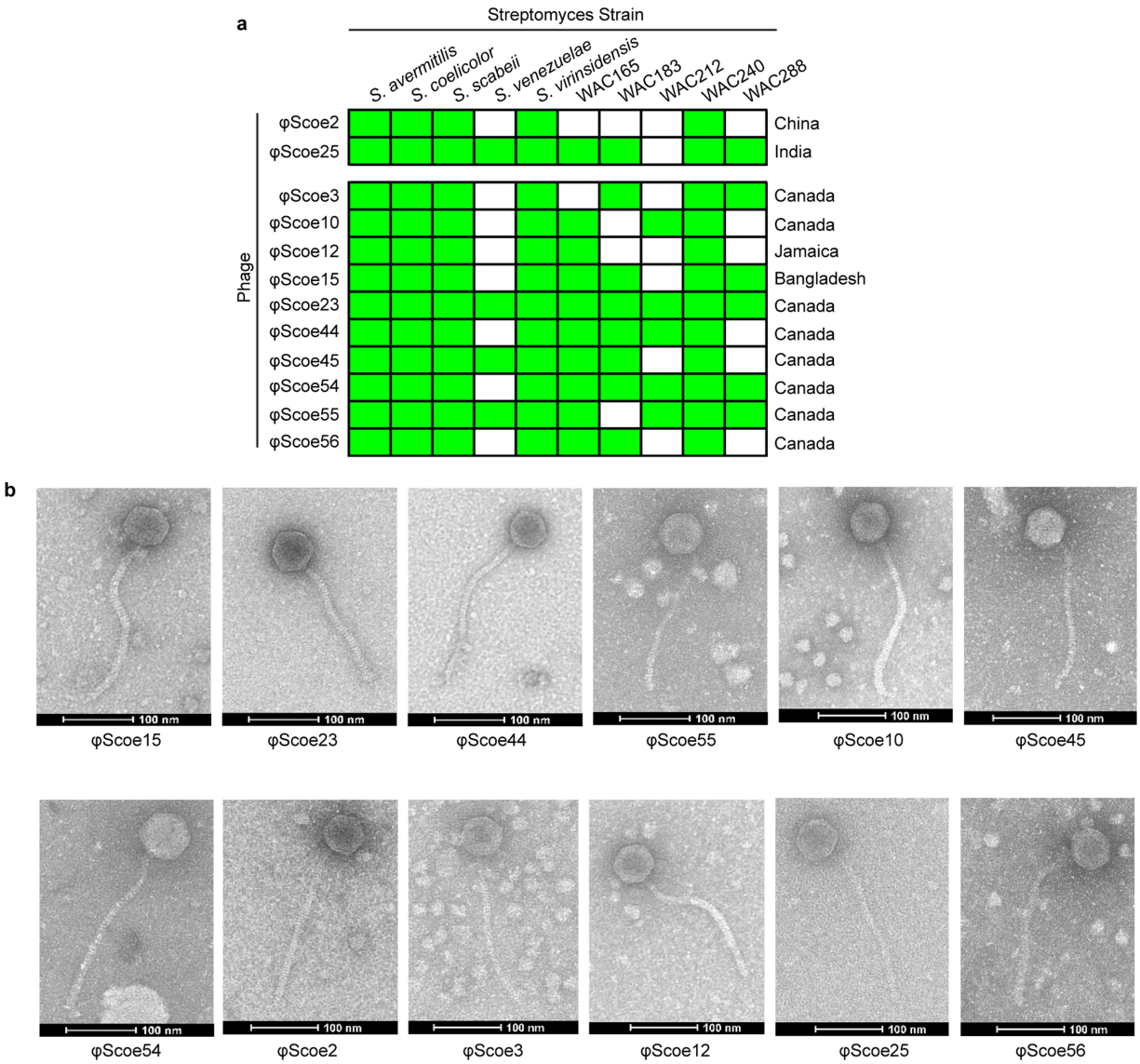
Induction of a temperature-sensitive lysogen. *E. coli* cells that contained a λcI_{857} prophage were grown overnight at 30 °C. Then, 1% subcultures were grown in the presence of 15 μM daunorubicin to $A_{600\,nm} = 0.5$. Cultures were transferred to a water bath shaker at 42 °C for 15 min to induce the temperature-sensitive lysogen. The cultures were grown at 37 °C for 2 h until complete cell lysis was observed. Phages were collected and active phage particles enumerated by spotting serial dilutions onto a lawn of *E. coli* BW25113 and counting plaques following overnight incubation at 37 °C.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

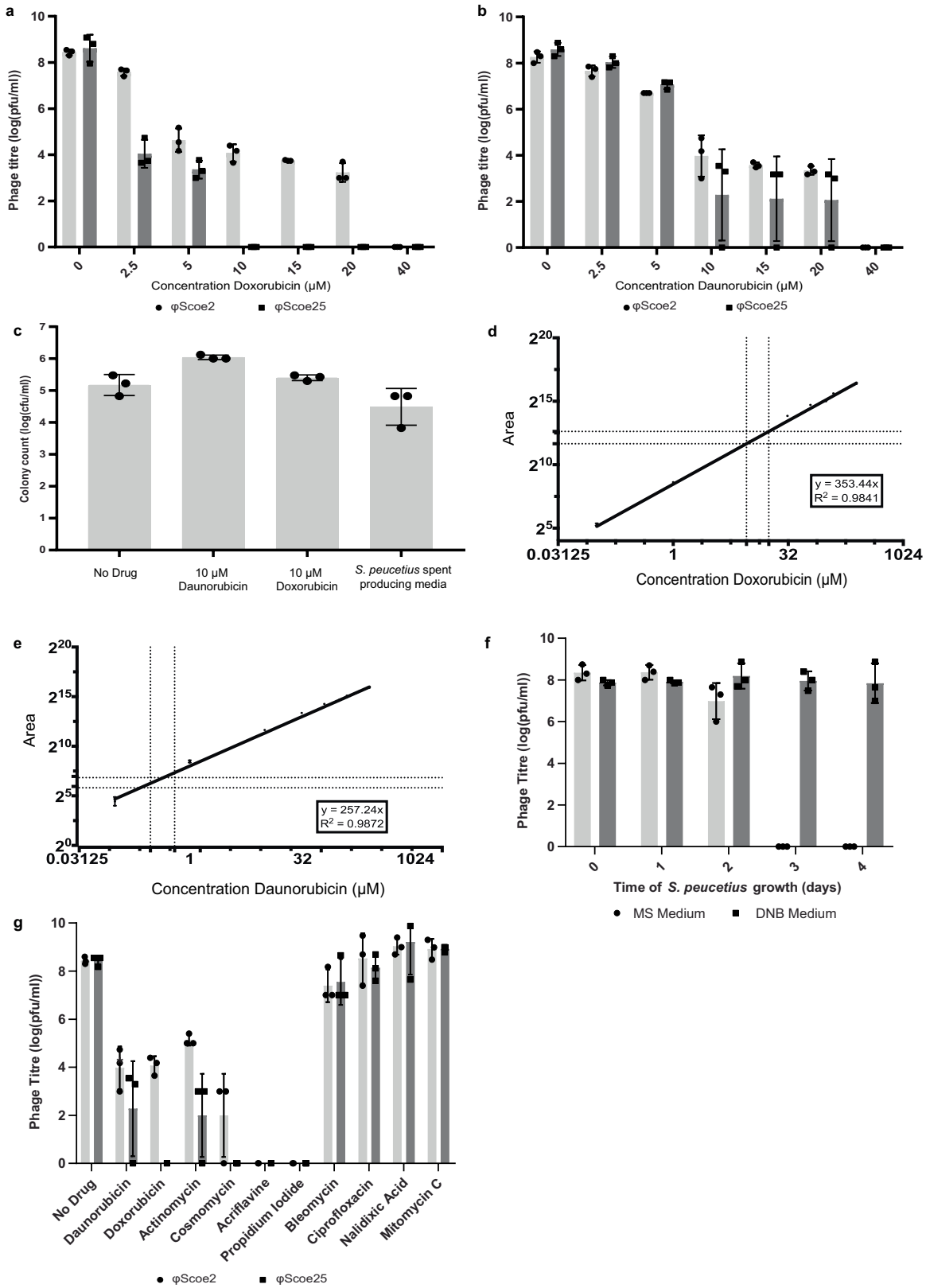
The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Characterization of *Streptomyces* phages used in this study. a, Host range profiles of the *Streptomyces* phages. *Streptomyces* phages were isolated from dirt samples from a number of different geographical locations as noted on the right. The host range of these phages was determined by plating on the panel of ten *Streptomyces* strains listed at the top ($n = 3$ biological replicates). Green boxes denote strains in which each phage was able to form plaques. Phages φScoe2 and φScoe25 were used in the initial screen of the 48 WAC extracts for anti-phage activity. **b**, Negatively stained phages were examined using TEM on a Talos L120C. Scale bars are indicatd at the bottom of each image. Each of the phages belongs to the *Siphoviridae* family, members of which have long non-contractile tails and an icosahedral head that contains a dsDNA genome. TEM grids were prepared once for each phage and two images were taken for each.

© 2018 Springer Nature Limited. All rights reserved.

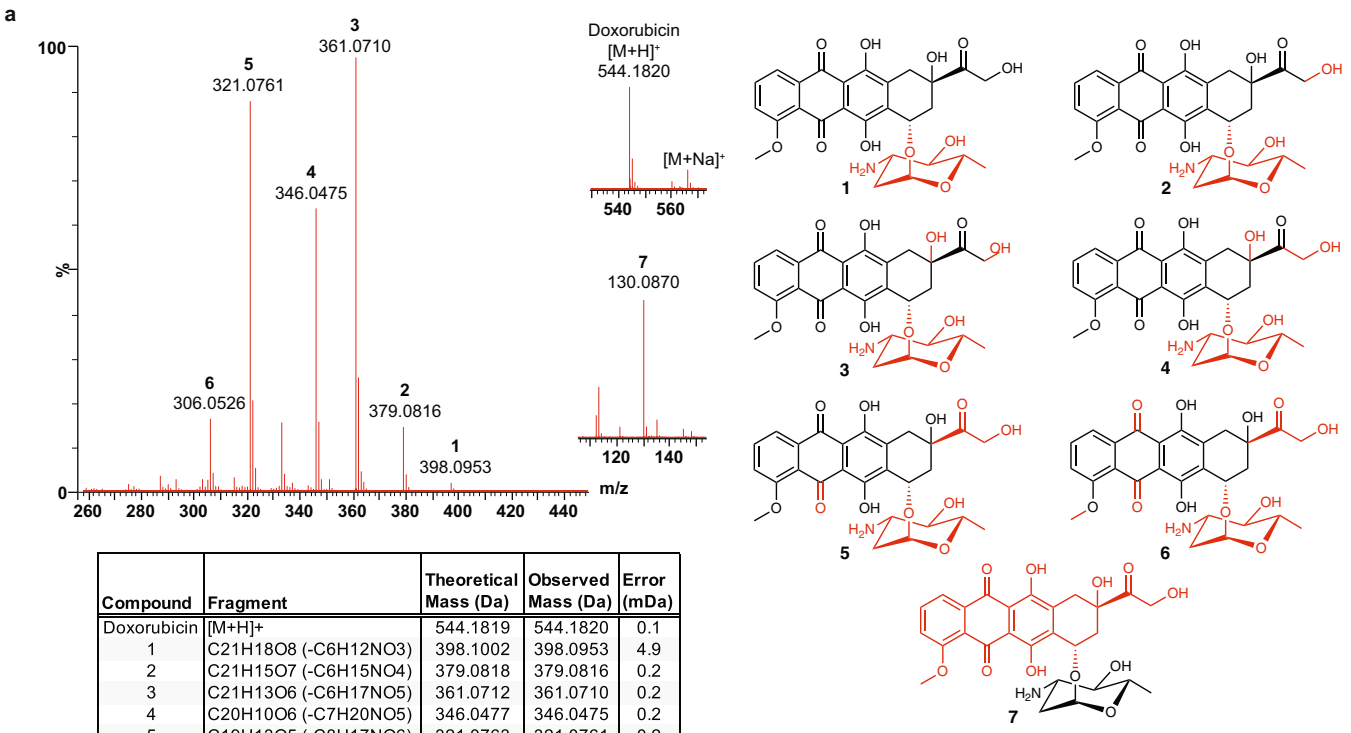


Extended Data Fig. 2 | See next page for caption.

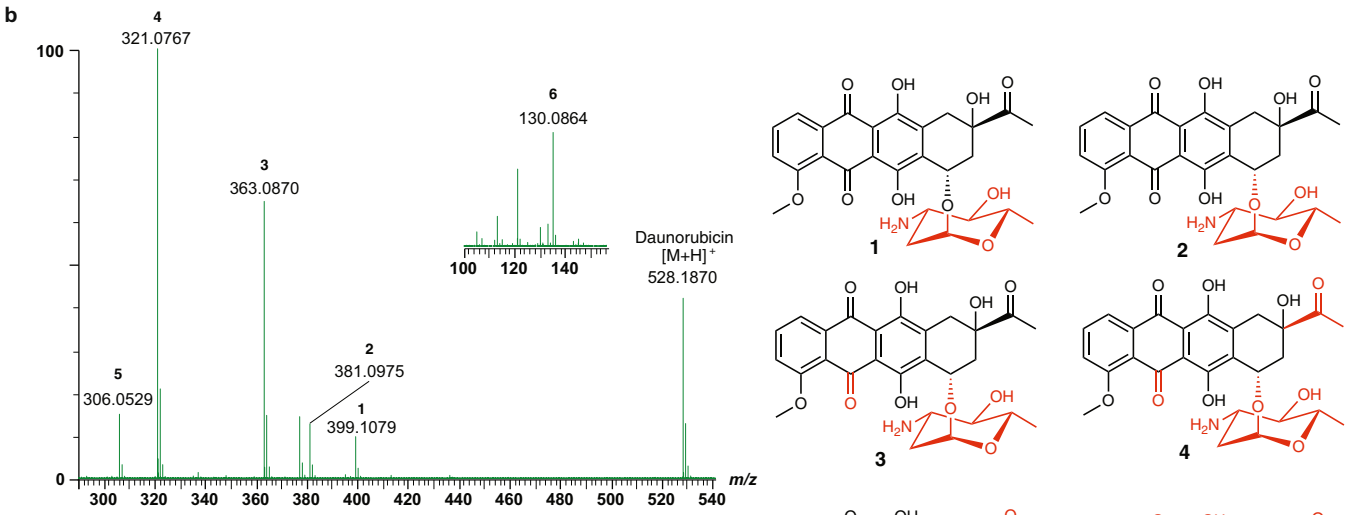
© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 2 | Effect of daunorubicin, doxorubicin and spent medium from *S. peucetius* on phage propagation and *S. coelicolor* growth. a, Determination of the minimum inhibitory concentration of doxorubicin on phages φScoe2 (black bars) and φScoe25 (grey bars) propagated in *S. coelicolor*. Phage titres were determined following propagation in *S. coelicolor* in the presence of doxorubicin at a range of concentrations that varied from 2.5 μM to 40 μM. Because 10 μM was the lowest concentration at which full inhibition of phage φScoe25 was observed, this was the baseline concentration selected for further experiments. Data are mean ±s.d.; $n = 3$ independent biological replicates. **b**, Determination of the minimum inhibitory concentration of daunorubicin on phages φScoe2 (black bars) and φScoe25 (grey bars) propagated in *S. coelicolor*. Phage titres were determined in *S. coelicolor* in the presence of daunorubicin at a range of concentrations that varied from 2.5 μM to 40 μM as in **a**. Data are mean ±s.d.; $n = 3$ independent biological replicates. **c**, Determination of the effects of 10 μM daunorubicin, 10 μM doxorubicin and spent medium from *S. peucetius* that contained doxorubicin and daunorubicin on the growth of *S. coelicolor*. The number of colony-forming units that are present after overnight growth of *S. coelicolor* in the presence of each of these compounds shows that they do not significantly decrease the growth of *S. coelicolor*. Data are mean ±s.d.; $n = 3$ independent biological replicates. **d**, Quantification of doxorubicin produced by *S. peucetius*. A standard curve was generated with commercially available doxorubicin by quantifying the ions that represent the proton adduct of the species ($n = 3$ independent experiments). Curves were then used to extrapolate the concentration of doxorubicin in *S. peucetius* culture supernatants after three and four days of growth. **e**, Quantification of daunorubicin produced by *S. peucetius*. A standard curve was generated with commercially available daunorubicin by quantifying the ions that represent the proton adduct of the species ($n = 3$ independent experiments). Curves were used to extrapolate the concentration of daunorubicin in *S. peucetius* culture supernatants after three and four days of growth. **f**, Effect of spent mannitol–soy and DNB media from *S. peucetius* on the propagation of phage φScoe2 after 1–4 days of *S. peucetius* growth. This graph reflects replicates of the phage spotting assay shown in Fig. 2b. Data are mean ±s.d.; $n = 3$ independent biological replicates. **g**, The effects of natural and synthetic compounds on the propagation of phages φScoe2 and φScoe25 were determined by overnight propagation of each phage in *S. coelicolor* in the presence of each compound at its specified concentration. This graph reflects replicates of the phage spotting assay shown in Fig. 2a. Data are mean ±s.d.; $n = 3$ independent biological replicates.

© 2018 Springer Nature Limited. All rights reserved.



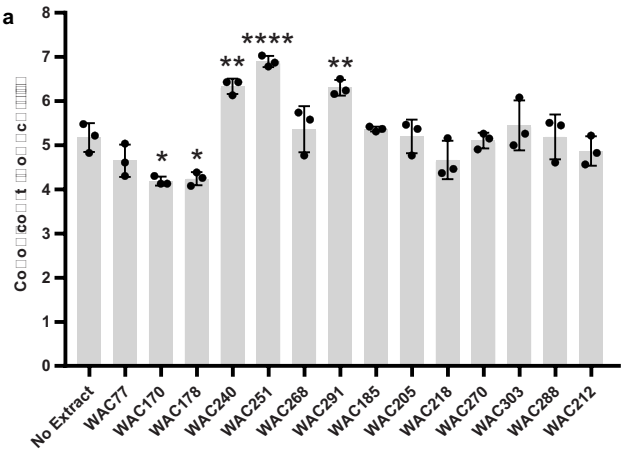
Compound	Fragment	Theoretical Mass (Da)	Observed Mass (Da)	Error (mDa)
Doxorubicin	[M+H]+	544.1819	544.1820	0.1
1	C21H18O8 (-C6H12NO3)	398.1002	398.0953	4.9
2	C21H15O7 (-C6H15NO4)	379.0818	379.0816	0.2
3	C21H13O6 (-C6H17NO5)	361.0712	361.0710	0.2
4	C20H10O6 (-C7H20NO5)	346.0477	346.0475	0.2
5	C19H13O5 (-C8H17NO6)	321.0763	321.0761	0.2
6	C19H13O4 (-C8H17NO7)	306.0528	306.0526	0.2
7	C6H12NO2 (-C21H18O9)	130.0868	130.0870	0.2

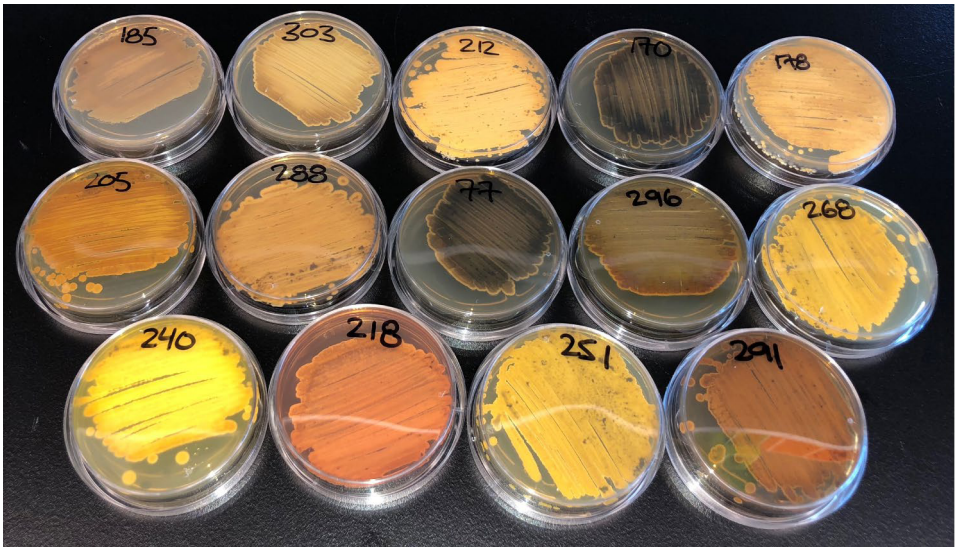


Compound	Fragment	Theoretical Mass (Da)	Observed Mass (Da)	Error (mDa)
Daunorubicin	[M+H]	528.1870	528.1870	0.0
1	$C_{21}H_{19}O_8$ (-$C_6H_{11}NO_2$)	399.1080	399.1079	-0.1
2	$C_{21}H_{17}O_7$ (-$C_6H_{13}NO_3$)	381.0974	381.0975	0.1
3	$C_{21}H_{15}O_6$ (-$C_6H_{15}NO_4$)	363.0869	363.0870	0.1
4	$C_{19}H_{13}O_5$ (-$C_8H_{17}NO_5$)	321.0763	321.0767	0.4
5	$C_{18}H_{10}O_5$ (-$C_9H_{20}NO_5$)	306.0528	306.0529	0.1
6	$C_6H_{10}NO$ (-$C_{21}H_{20}O_9$)	130.0868	130.0864	-0.4

Extended Data Fig. 3 | Detection of doxorubicin and daunorubicin in spent medium from *S. peucetius*. **a**, Doxorubicin production in mannitol–soy broth was confirmed by ultra-performance liquid chromatography–tandem mass spectrometry with comparison to a doxorubicin standard. The observed fragmentation from a sample grown in doxorubicin-producing permissive medium adhered to the doxorubicin structure with minimal mass errors. Complementary fragments (**1** and **6**) with their losses highlighted in red further support the doxorubicin structure. In addition, the fragmentation observed from extracts matches a doxorubicin commercial standard ($n = 1$ experiment). **b**, Daunorubicin production in mannitol–soy broth was confirmed by ultra-performance liquid chromatography–tandem mass spectrometry with comparison to a daunorubicin standard. The observed fragmentation from a sample grown in daunorubicin-producing permissive medium adhered to the daunorubicin structure with minimal mass errors. Complementary fragments (**1** and **6**) with their losses highlighted in red further support the daunorubicin structure. In addition, the fragmentation observed from extracts matches a daunorubicin commercial standard ($n = 1$ experiment).

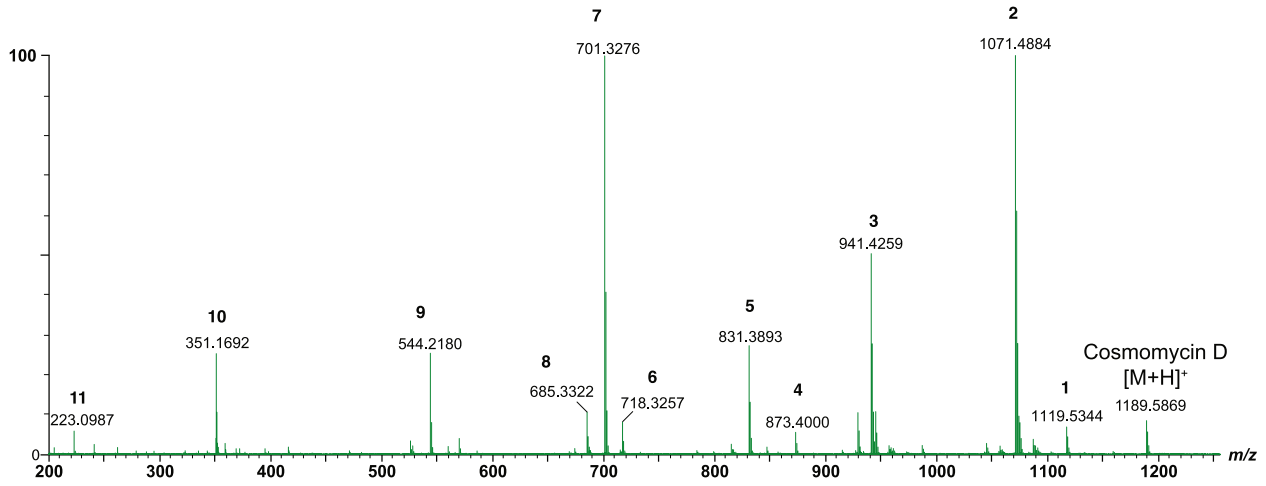
© 2018 Springer Nature Limited. All rights reserved.

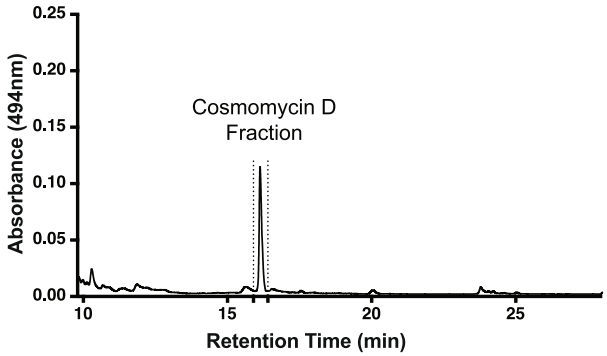




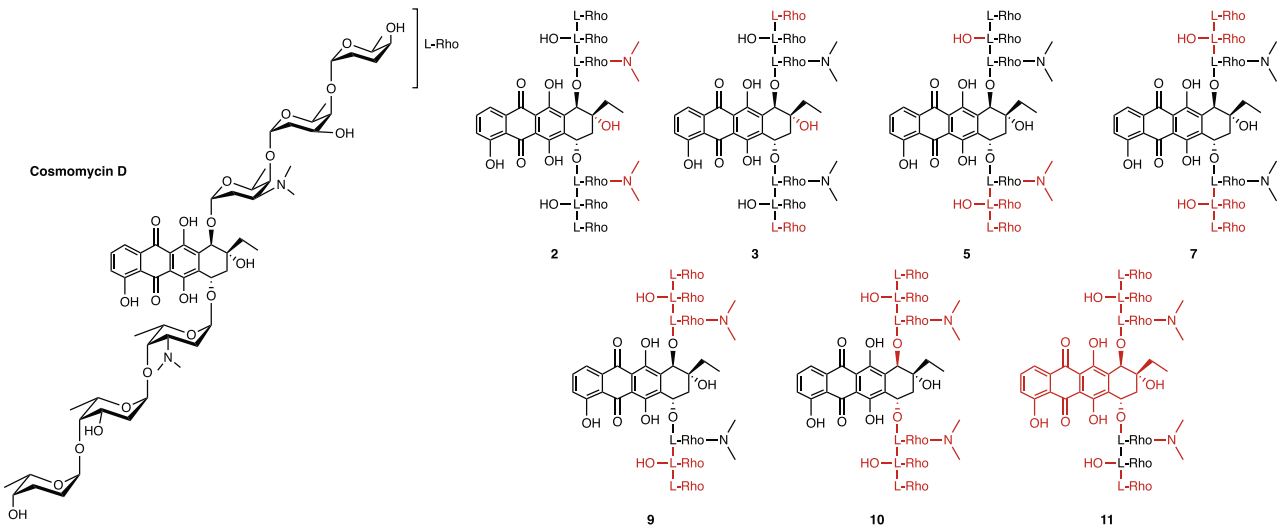
Extended Data Fig. 4 | Effect of active WAC extracts on the growth of *S. coelicolor* and diversity of secondary metabolites produced. **a**, *S. coelicolor* was incubated overnight in the presence of each of the fourteen WAC extracts that showed anti-phage activity. The majority of extracts had no effect on bacterial cell growth. Two extracts, WAC170 and WAC178, inhibited growth of *S. coelicolor* approximately tenfold. Data are mean ±s.d.; $n = 3$ independent biological replicates. Statistically different samples compared to the untreated sample by ANOVA are indicated (from left to right, $^{*}P = 0.0117$, $^{*}P = 0.0188$, $^{**}P = 0.0022$, $^{****}P < 0.0001$, $^{**}P = 0.0029$). **b**, Growth of the 14 WAC strains for which anti-phage activity was detected on MYM solid medium reveals the diversity of secondary metabolite pigment production ($n = 3$ biologically independent replicates).

© 2018 Springer Nature Limited. All rights reserved.



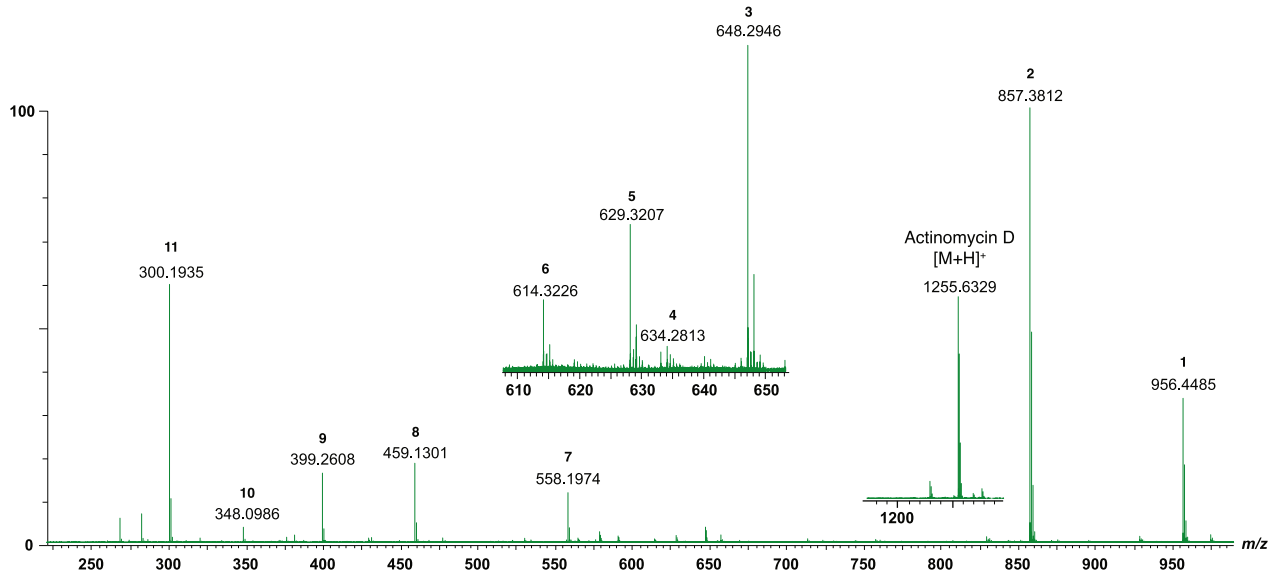


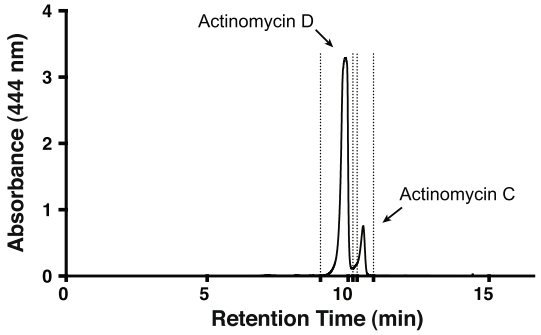
Compound	Fragment	Theoretical Mass (Da)	Observed Mass (Da)	Error (mDa)
Cosmomycin D	[M+H]	1189.5907	1189.5869	-3.8
1	$C_{59}H_{79}N_2O_{19}$ (-$CH_{10}O_3$)	1119.5277	1119.5344	6.7
2	$C_{55}H_{75}O_{21}$ (-$C_5H_{14}N_2O$)	1071.4801	1071.4884	8.3
3	$C_{48}H_{65}N_2O_{17}$ (-$C_{12}H_{24}O_5$)	941.4283	941.4259	-2.4
4	$C_{44}H_{61}N_2O_{16}$ (-$C_{16}H_{28}O_6$)	873.4021	873.4000	-2.1
5	$C_{42}H_{59}N_2O_{15}$ (-$C_{18}H_{30}O_7$)	831.3915	831.3893	-2.2
6	$C_{40}H_{48}NO_{11}$ (-$C_{20}H_{41}NO_{11}$)	718.3227	718.3257	3.0
7	$C_{36}H_{49}N_2O_{12}$ (-$C_{24}H_{40}O_{10}$)	701.3285	701.3276	-0.9
8	$C_{36}H_{49}N_2O_{11}$ (-$C_{24}H_{40}O_{11}$)	685.3336	685.3322	-1.4
9	$C_{28}H_{34}NO_{10}$ (-$C_{32}H_{55}NO_{12}$)	544.2183	544.2180	-0.3
10	$C_{20}H_{15}O_6$ (-$C_{40}H_{74}N_2O_{16}$)	351.0869	351.0876	0.7
11	$C_{12}H_{15}O_4$ (-$C_{48}H_{74}N_2O_{18}$)	223.0970	223.0987	1.7



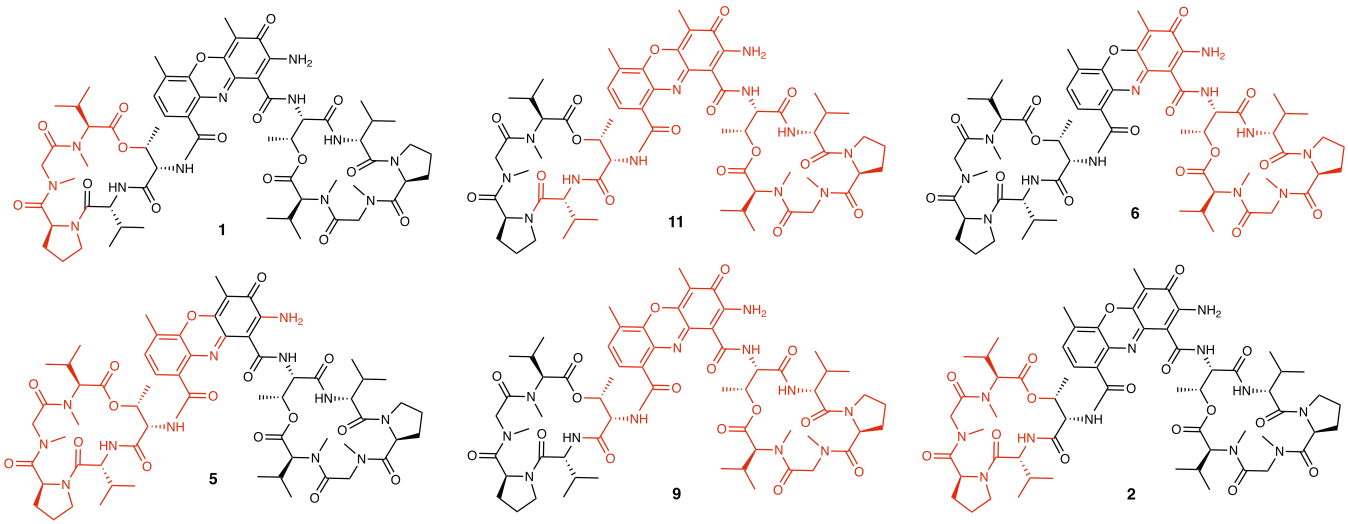
Extended Data Fig. 5 | Elucidation of active secondary metabolites in WAC288. Cosmomycin D was determined to be the active metabolite within the WAC288 extract following a bioactivity-guided fractionation strategy ($n = 3$ biological replicates). The chromatogram ($A_{494\,nm}$) of the HPLC purification indicates the cosmomycin D fraction collected for bioactivity assays and structure elucidation. The cosmomycin D structure was confirmed by high-resolution tandem mass spectrometry of its associated proton adduct (1,189.5869 *m/z*). The observed fragmentation supports the cosmomycin D structure with minimal mass errors for each fragment ($n = 1$ experiment). Key fragment losses are annotated with their associated peak, and their losses are highlighted in red.

© 2018 Springer Nature Limited. All rights reserved.



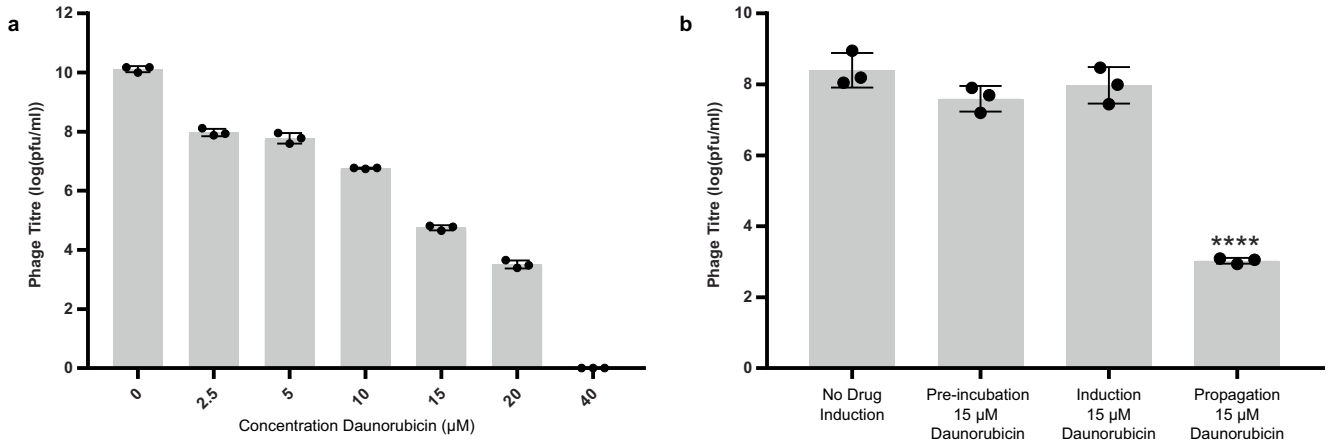


Compound	Fragment	Theoretical Mass (Da)	Observed Mass (Da)	Error (mDa)
Actinomycin D	[M+H]	1255.6363	1255.6329	-3.4
1	$C_{48}H_{62}N_9O_{12}$ (-$C_{14}H_{25}N_3O_4$)	956.4518	956.4485	-3.3
2	$C_{43}H_{52}N_8O_{11}$ (-$C_{19}H_{34}N_4O_5$)	857.3834	857.3812	-2.2
3	$C_{33}H_{40}N_6O_8$ (-$C_{29}H_{47}N_6O_8$)	648.2908	648.2946	3.8
4	$C_{31}H_{44}N_6O_8$ (-$C_{31}H_{43}N_6O_8$)	634.2751	634.2813	6.2
5	$C_{31}H_{45}N_6O_8$ (-$C_{31}H_{42}N_6O_8$)	629.3299	629.3207	-9.2
6	$C_{30}H_{59}N_9O_8$ (-$C_{32}H_{59}N_9O_8$)	558.1876	558.1974	9.8
7	$C_{24}H_{19}N_4O_6$ (-$C_{38}H_{68}N_8O_{10}$)	459.1305	459.1301	-0.4
8	$C_{19}H_{35}N_4O_5$ (-$C_{43}H_{52}N_8O_{11}$)	399.2607	399.2608	0.1
9	$C_{19}H_{14}N_3O_4$ (-$C_{43}H_{14}N_9O_{12}$)	348.0984	348.0986	0.2
10	$C_{14}H_{26}N_3O_4$ (-$C_{48}H_{61}N_9O_{12}$)	300.1923	300.1935	1.2



Extended Data Fig. 6 | Elucidation of active secondary metabolites in WAC240. The active molecule in the WAC240 extract was determined by a bioactivity-guided fractionation approach ($n = 3$ biological replicates). The chromatogram ($A_{444\,nm}$) of the final HPLC purification step highlights the actinomycin D fraction collected. Structure elucidation was accomplished via high-resolution tandem mass spectrometry of the actinomycin D proton adduct (1255.6329 m/z) following HPLC purification ($n = 1$ experiment). All of the fragments support the actinomycin D structure with minimal mass errors, and the key fragment losses are highlighted in the structures in red. Many of the annotated fragments are complementary (for example, **1** and **11**), further supporting that the parent ion is actinomycin D.

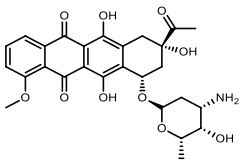
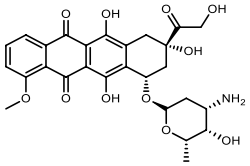
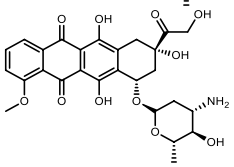
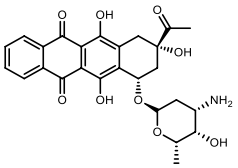
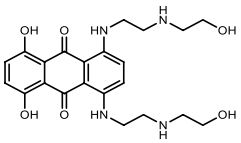
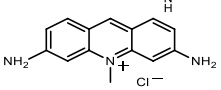
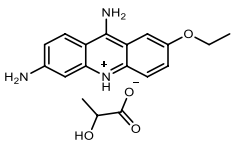
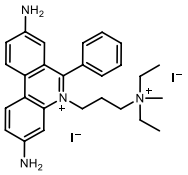
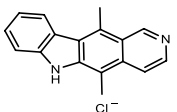
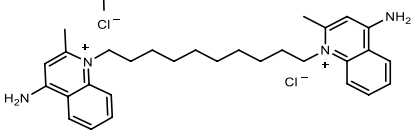
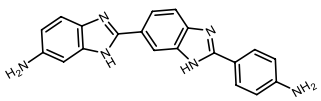
© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Effect of daunorubicin on the life cycle of the *E. coli* phage λ. a, Determination of the working concentration of daunorubicin for *E. coli* phage λ. Phage λ was propagated in the presence of daunorubicin at concentrations that varied from 2.5 μM to 40 μM and the resulting phage titres were determined by spotting serial dilutions of the lysate onto a lawn of *E. coli* BW25113. Because daunorubicin at 15 μM provided a 10^5-fold decrease in phage titre and was similar in concentration to the 10-μM concentration used in the *Streptomyces* phage assays, it was chosen as the working concentration for phage λ. Data are mean ±s.d.; $n = 3$ independent biological replicates. **b**, Phage λ was propagated in *E. coli* BW25113 for 6 h in the absence and presence of 15 μM daunorubicin and the resulting phages were enumerated by plating serial dilutions of the supernatant on BW25113. When the phage was pre-incubated with daunorubicin, there was no significant decrease in phage titre. Induction of a temperature-sensitive λ lysogen in the presence of daunorubicin also did not decrease the phage titre. This graph reflects replicates of the phage spotting assay shown in Fig. 3c. Data are mean ±s.d.; $n = 3$ biological replicates. The statistically different sample compared to the untreated sample by ANOVA is indicated (****$P < 0.0001$).

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Compounds that are shown to inhibit propagation of *E. coli* phage λ

Compound	Mode of Action	Structure
Daunorubicin	DNA intercalation	
Doxorubicin	DNA intercalation	
Epirubicin	DNA intercalation	
Idarubicin	DNA intercalation	
Mitoxantrone	DNA intercalation	
Acriflavine	DNA intercalation	
Ethacridine lactate	DNA intercalation	
Propidium iodide	DNA intercalation	
Ellipticine	DNA intercalation	
Dequalinium chloride	Disrupts membrane permeability	
Ro 90-7501	Inhibits Ser/Thr phosphatase	

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Analytical profile index 20E biochemical characterization of WAC strains

		Wild Isolates													
		77	170	178	185	205	212	218	240	251	268	288	291	296	303
	Genus - 16S Sequencing	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*	*Strept.*
	Sampling Location	New Westminster, British Columbia, Canada	West Vancouver, British Columbia, Canada	Hamilton, Ontario, Canada	Regina, Saskatchewan, Canada	Hamilton, Ontario, Canada	Hamilton, Ontario, Canada	Mississauga, Ontario, Canada	Surrey, British Columbia, Canada	New Westminster, British Columbia, Canada	New Westminster, British Columbia, Canada	Ancaster, Ontario, Canada	Ancaster, Ontario, Canada	Burlington, Ontario, Canada	Ancaster, Ontario, Canada
Test	**Description**														
ONPG	β-galactosidase activity	-	-	+	-	+	-	-	+	-	-	+	+	+	-
ADH	Arginine dihydrolase activity	+	+	-	-	+	+	-	+	-	-	+	+	+	+
LDC	Lysine decarboxylase activity	-	-	-	-	-	-	-	+	-	-	-	-	-	-
ODC	Ornithine decarboxylase activity	-	-	-	-	-	-	-	+	-	-	-	-	-	-
CIT	Citrate utilization	-	-	-	-	+	-	-	-	-	-	-	-	-	-
H2S	Hydrogen sulfide production	+	+	-	-	-	-	-	-	-	-	-	-	+	-
URE	Urease activity	+	-	-	-	+	+	-	+	+	+	+	-	-	+
TDA	Tryptophan deaminase activity	+	+	-	-	-	-	-	+	-	-	+	+	+	-
IND	Indol production	-	-	-	-	-	-	-	-	-	-	-	-	-	-
VP	Acetyl methylcarbinol production	+	+	+	+	+	+	+	+	+	+	+	-	+	+
GEL	Gelatinase activity	+	+	+	+	+	+	+	+	+	+	+	+	+	+
GLU	Glucose fermentation	+	+	-	+	+	+	+	+	+	+	+	-	-	+
MAN	Mannose fermentation	-	-	+	+	+	+	-	+	+	-	-	+	-	+
INO	Inositol fermentation	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SOR	Sorbitol fermentation	-	+	-	-	-	+	-	-	-	-	-	+	-	-
RHA	Rhamnose fermentation	-	+	-	-	+	-	+	-	+	+	-	-	-	+
SAC	Sucrose fermentation	-	-	-	-	-	-	+	-	-	-	-	-	-	-
MEL	Melibiose fermentation	-	-	-	-	-	-	-	-	-	-	+	-	-	-
AMY	Amygdalin fermentation	-	+	+	-	+	+	+	+	+	+	+	-	+	+
ARA	Arabinose fermentation	+	+	-	+	+	+	+	+	+	+	+	+	+	+

+ = postive - = negative

Phenotypes were determined for each isolate from the WAC that was identified as producing an anti-phage compound (see Fig. 2c). Assays were performed using the analytical profile index 20E test by bioMerieux, according to their protocol.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0769-8

Dna2 nuclease deficiency results in large and complex DNA insertions at chromosomal breaks

Yang Yu[1,6], Nhung Pham[1,6], Bo Xia[2,3,4,6], Alma Papusha[1], Guangyu Wang[2,3,4], Zhenxin Yan[1], Guang Peng[5], Kaifu Chen[2,3,4]* & Grzegorz Ira[1]*

Insertions of mobile elements[1–4], mitochondrial DNA[5] and fragments of nuclear chromosomes[6] at DNA double-strand breaks (DSBs) threaten genome integrity and are common in cancer[7–9]. Insertions of chromosome fragments at V(D)J recombination loci can stimulate antibody diversification[10]. The origin of insertions of chromosomal fragments and the mechanisms that prevent such insertions remain unknown. Here we reveal a yeast mutant, lacking evolutionarily conserved Dna2 nuclease, that shows frequent insertions of sequences between approximately 0.1 and 1.5 kb in length into DSBs, with many insertions involving multiple joined DNA fragments. Sequencing of around 500 DNA inserts reveals that they originate from Ty retrotransposons (8%), ribosomal DNA (rDNA) (15%) and from throughout the genome, with preference for fragile regions such as origins of replication, R-loops, centromeres, telomeres or replication fork barriers. Inserted fragments are not lost from their original loci and therefore represent duplications. These duplications depend on nonhomologous end-joining (NHEJ) and Pol4. We propose a model in which alternative processing of DNA structures arising in Dna2-deficient cells can result in the release of DNA fragments and their capture at DSBs. Similar DNA insertions at DSBs are expected to occur in any cells with linear extrachromosomal DNA fragments.

We analysed DSB repair by NHEJ in yeast cells deficient in the nuclease/helicase Dna2 and found that approximately 8% of the repair events carried large insertions of about 100–1,500 bp, whereas the remaining repair events were comparable to those found in wild-type cells (Fig. 1a, b, Extended Data Fig. 1, Extended Data Tables 1, 2, Supplementary Information Table 1). In this experimental design[11], homothalic switching (HO) endonuclease-induced DSBs at the *MAT**a*** locus can only be repaired by imprecise NHEJ that alters the HO-cleavage site, preventing further cutting. Analysis was done in cells carrying a suppressor of *dna2*Δ lethality, the *pif1-m2* mutation[12]. No insertions were found in *pif1-m2* or wild-type control cells. The nuclease activity of Dna2, but not its helicase activity, is required to suppress insertions at DSBs (Fig. 1a). We observed similar insertions to those found at the *MAT* locus at a DSB induced at an artificially introduced *ACT1* intron within the *URA3*[3] locus or at CRISPR–Cas9-induced DSBs in the *LYS2* gene (Extended Data Fig. 1, Extended Data Table 1). Sequencing analysis of about 500 insertions from all Dna2-deficient cells reveals that approximately 15% of events contained 2 to 4 fragments from different chromosomes joined together at the DSB (Fig. 1c, d).

The DNA insertions result in duplications, as none of 25 randomly tested donor DNA fragments were deleted from their original locus, and the number of insertions originating from essential genes (46/222 in all strains tested) was proportional to the number of essential genes in yeast (~20%). The duplicated sequences include short complete genes, replication origins, and fragments of telomeres or centromeres (Supplementary Information Table 1). NHEJ is the primary pathway mediating these insertions, as most of the junctions carried 0–4 nucleotides of microhomology (Fig. 1e), DSB ends were mostly maintained (Extended Data Fig. 1) and deletion of NHEJ components (Ku, Lig4 or Pol4) nearly abolished insertions (Fig. 1a). Single insertion captured in NHEJ-deficient cells shows an increased microhomology and loss of sequences at DSB ends, typical features of alternative end joining. By contrast, deletion of homologous-recombination-specific enzymes in *pif1-m2 dna2*Δ cells had no effect on the number of insertions in the case of Rad51, and increased the number of insertions in the case of Rad52 (Fig. 1a).

The origins of inserted DNA in *pif1-m2 dna2*Δ can be grouped into four major categories. First, about 8% of the insertions are fragments of retrotransposons, which comprise about 3% of the 12.1-Mb yeast genome. Second, about 15% of insertions originate from rDNA,

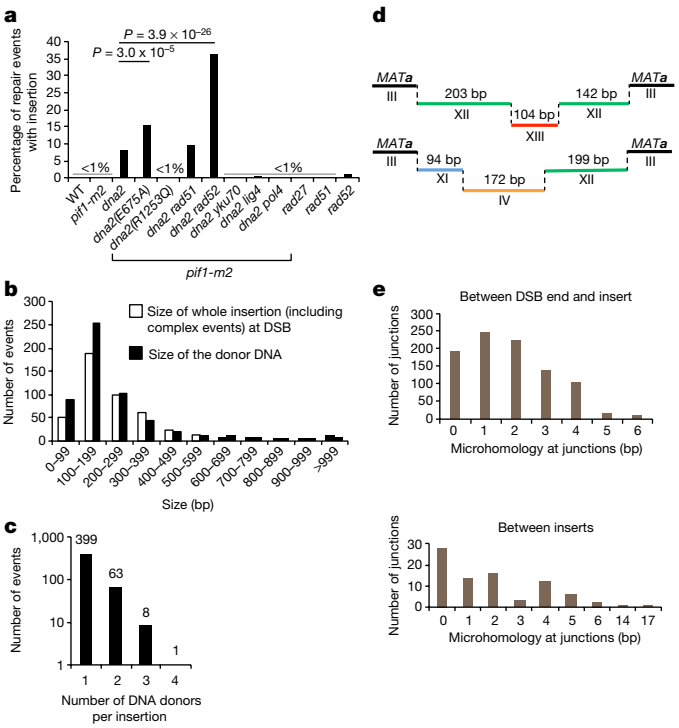


Fig. 1 | Dna2 inhibits large insertions at DSBs. a, Frequency of insertions in indicated mutants. χ^2 test is used to determine the *P* value; number of colonies tested per mutant is indicated in Extended Data Table 1. WT, wild type. *lig4* is also known as *dnl4*. **b**, Insertion size analysis in Dna2-deficient cells. **c**, Number of DNA fragments per insertion. **d**, Two examples of complex insertions at DSBs. Roman numerals indicate the chromosome of origin of the sequence. **e**, Analysis of microhomologies at junctions of inserted DNA or between inserted fragments.

[1]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, USA. [2]Center for Bioinformatics and Computational Biology, Department of Cardiovascular Sciences, Houston Methodist Research Institute, Houston, TX, USA. [3]Institute for Academic Medicine, Houston Methodist Hospital, Houston, TX, USA. [4]Weill Cornell Medical College, Cornell University, Department of Cardiothoracic Surgery, New York, NY, USA. [5]Department of Clinical Cancer Prevention, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [6]These authors contributed equally: Yang Yu, Nhung Pham, Bo Xia. *e-mail: kchen2@houstonmethodist.org; gira@bcm.edu

© 2018 Springer Nature Limited. All rights reserved.

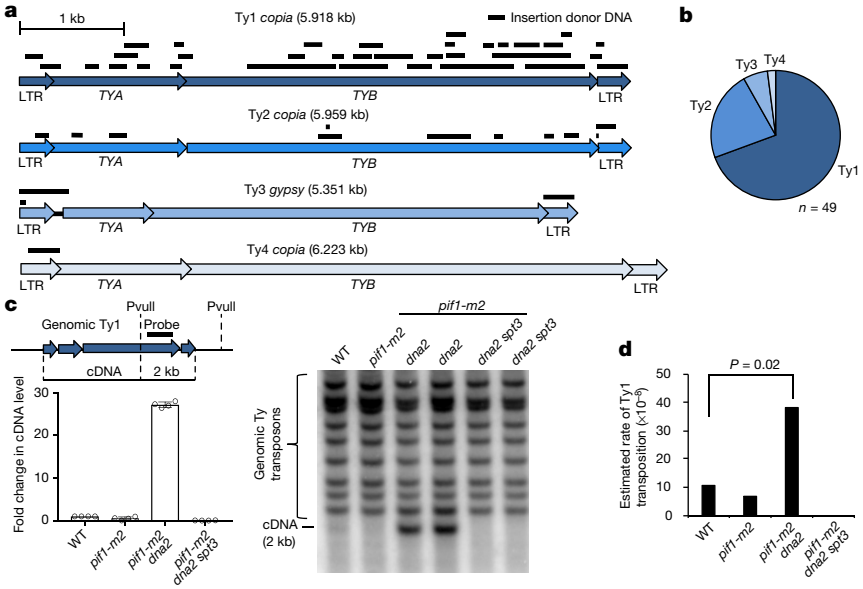


Fig. 2 | Dna2 limits mobility of transposable elements. a, Position of inserted DNA originating from yeast transposons. LTR, long terminal repeat. **b**, Contribution of yeast transposons to insertions. **c**, Scheme of Ty1 cDNA analysis, Southern blot analysis and quantification of Ty1 cDNA in indicated mutants. Data are mean ± s.d.; $n = 3$ independent experiments. For gel source data, see Supplementary Figure 1. **d**, Retrotransposition rates in indicated mutants. Bootstrap resampling (two-tailed) to determine P value; $n = 4$ independent experiments.

representing about 10% of the genome. Third, about 74% of insertions originate from elsewhere in the genome. Finally, about 3% of insertions originate from 6.3-kb resident 2μ DNA plasmids (approximately 50 copies per cell), reflecting the proportion of these plasmids in nuclear DNA content. Mitochondrial DNA was not inserted. Proximity of insertion donor DNA to the DSB is not important, as few insertions originate from the chromosome carrying the DSB, and the number of donor DNAs from different chromosomes correlates simply with their size, with the exception of the rDNA cluster on chromosome XII, which contains the largest hotspot of insertion donor DNA (Extended Data Fig. 2). Accordingly, analysis of the three-dimensional proximity of the donor DNAs or randomly selected sequences to the locus of DSBs, as measured by chromosome conformation capture[13], were not distinguishable (Extended Data Fig. 2).

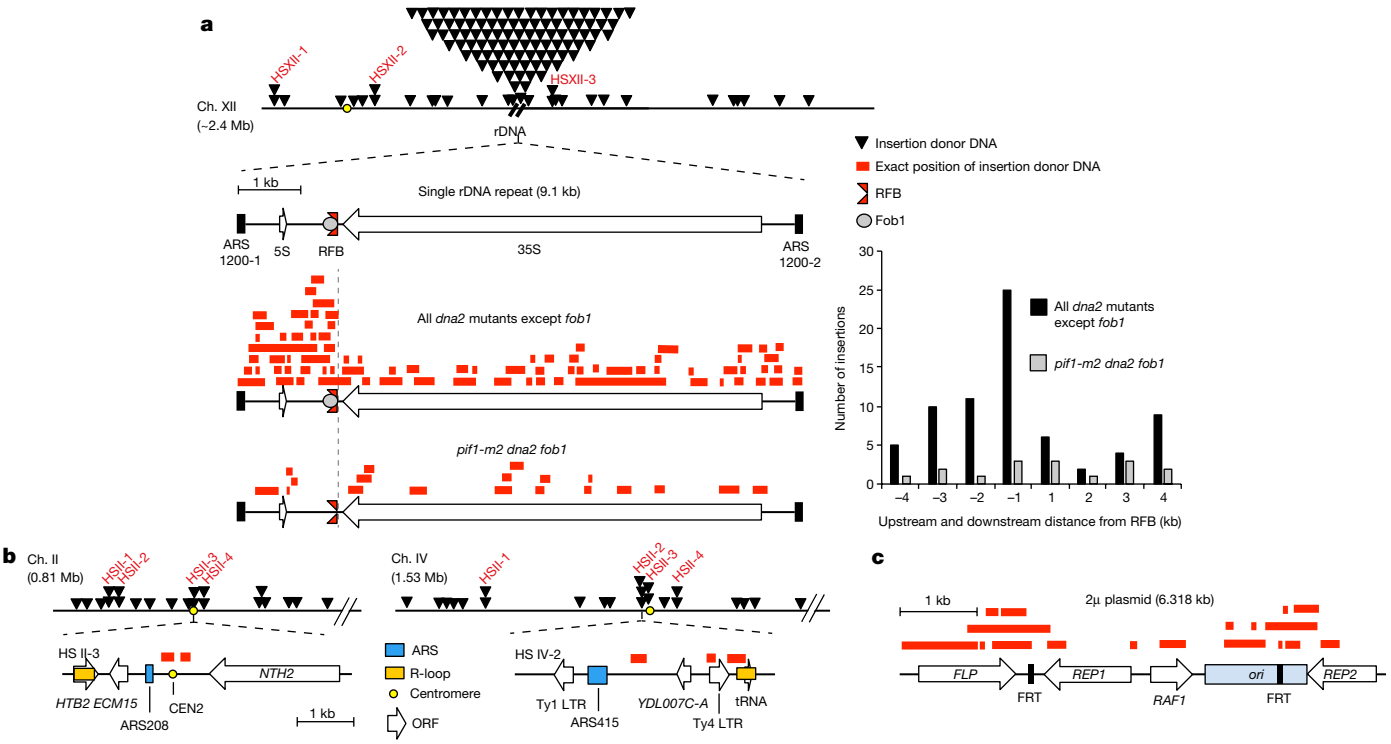


Fig. 3 | Origin of inserted DNA at DSBs. a, Top, location of inserted DNA at DSBs originating from chromosome (ch.) XII. Hotspots (HS, numbered), loci of up to 3 kb that provide at least two inserted fragments, are marked in red. Schematic of single rDNA repeat (second from top) and position of inserted DNA originating from rDNA repeats (bottom two). Right, number of insertions with respect to RFB position. **b**, Examples of hotspots of the origin of insertions with genomic features shown. ORF, open reading frame. **c**, Inserted DNA originating from 2μ plasmid. FRT, Flp recognition target; *ori*, origin of replication sequence.

© 2018 Springer Nature Limited. All rights reserved.

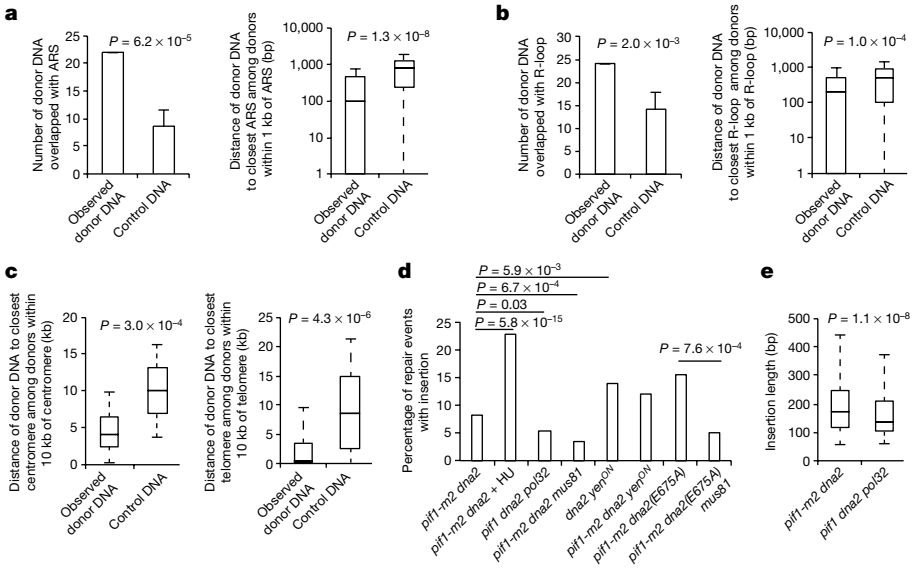


Fig. 4 | Insertions originate from regions where replication forks stall. a–**c**, Plots showing overlap or proximity of observed insertions and control DNA with indicated genomic features. *P* values for overlap and proximity are determined by bootstrapping and one-tailed Wilcoxon test, respectively. **a**, **b**, $n = 370$. **c**, $n = 370$ for centromere proximity analysis, $n = 371$ for telomere proximity analysis; *n* represents the number of independent insertions. Experiments with control DNA were repeated 1,000,000 times (**a**) or 1,000 times (**b**, **c**). **d**, Insertion frequencies in indicated mutants. χ^2 test is used to determine *P* values; number of colonies tested per mutant is indicated in Extended Data Table 1. HU, hydroxyurea. **e**, Insertion length analysis in indicated mutants. *P* value is calculated by one-tailed, one-sample Wilcoxon test; number of independent insertions per mutant is shown in Extended Data Table 1. Bar graphs: data are mean ± s.d. Box plots: centre line is median, boxes show first and third quartiles, whiskers extend to the most extreme data points that are no more than 1.5 fold of the interquartile range from the box.

All four active yeast transposons were present among insertions at DSBs (Fig. 2a, b), while only the most abundant Ty1 transposon insertions have previously been reported[1,3]. Deletion of Spt3—which is required for transposon transcription[14]—in *pif1-m2 dna2*Δ cells decreases the number of transposon insertions (from 8%) to the proportion of transposons in the genome (approximately 3%) (Extended Data Table 1), indicating that the reverse transcriptase activity of retrotransposons is important for transposon insertion. Ty1 cDNA levels (Fig. 2c) and the rate of retrotransposition measured via Ty1–His3 reporter[15] also increased with *DNA2* deletion (Fig. 2d). Increased levels of cDNA in *pif1-m2 dna2*Δ are not related to increased transcription of transposons, but may result from increased cDNA stability (Extended Data Fig. 3). Together, Dna2 inhibits retrotransposition and insertions of transposon fragments at DSBs.

About 15% of insertions originate from approximately 150 rDNA repeats. Each 9.1-kb repeat contains 5S and 35S genes, origin of replication (ARS) and replication fork blocking (RFB) sequences (Fig. 3a). Binding of Fob1 protein to RFB sequences prevents head-on collisions between replication forks and 35S transcription bubbles. Most rDNA inserted at DSBs originate from the region between ARS and RFB sequences, in which Dna2 prevents fork stalling[16]. This distribution is dependent on Fob1 and, therefore, on fork pausing at RFB (Fig. 3a). Out of 41 donor DNA hotspots providing at least 2 inserted fragments from within a 3-kb region, 34 are located in the vicinity of an ARS. Further, nearly half of 18 insertions from 2μ plasmids come from replication origins (Fig. 3b, c, Extended Data Fig. 4). Genome-wide analysis of the overlap or proximity of insertion donor DNAs to ARS sequences confirms this correlation (Fig. 4a). Donor DNAs were found to be nearer to sites of prominent R-loops, centromeres or telomeres, when compared to randomly selected sequences of equal size and frequency per chromosome (Fig. 4b, c). These features are known to cause fork stalling[17–19], and require Dna2 for replication to occur correctly (for example, refs [20,21]). Finally, treatment of *pif1-m2 dna2*Δ cells with a high dose of hydroxyurea, a drug known to cause fork stalling and reversal[22], results in an approximately twofold increase in insertions events (Fig. 4d). Together, these results show that inserted donor DNAs often originate from fragile genomic regions in which fork stalling is more likely.

As inserted DNA is not deleted from its original locus, it must either be over-replicated and then inserted into a DSB, or originate from a fragmented sister chromatid. We favour the first of these scenarios, because Dna2 has two functions that prevent over-replication: it removes long 5′ flaps during lagging-strand synthesis, and it prevents and/or degrades reversed forks[22–24]. Long, unprocessed 5′ flaps may contribute to insertions, because deletion of Pol32—the processivity subunit of Polδ—and Pif1 helicase, both of which stimulate the displacement synthesis that generates long flaps[25,26], reduces insertion frequency by about 50% and decreases the mean size of the insertions (Fig. 4d, e). Moreover, the size range of the insertions observed in these experiments resembles that seen for 5′ flaps in Dna2-deficient mutants[24]. Overexpression of Rad52 was previously shown to reduce the level of Dna2 substrates, presumably 5′ flaps[27]. Consistent with these results, we found a marked increase in number of insertions in *pif1-m2 dna2*Δ *rad52*Δ cells. We note that *rad52*Δ cells contained rare insertions (1%) of DNA from a 2-kb region on either side of the DSB (Extended Data Fig. 5). Deletion of the nonessential Rad27, which processes much shorter 5′ flaps, does not result in insertions, suggesting that there is efficient alternative processing of the flaps in *rad27*Δ cells (Fig. 1a). Reversed forks could also contribute, because inserted DNAs originate from genomic regions that are prone to fork stalling (Fig. 4a–c). In the absence of Dna2, unprocessed DNA structures can be cleaved by alternative nucleases, leading to release of DNA fragments that could be subsequently inserted into DSBs. A significant increase of insertions in $dna2^{E675A}$ cells suggests that the nuclease-dead Dna2 may bind and stabilize such DNA structures (Fig. 1a). Deletion of structure-specific nuclease Mus81, which cleaves stalled or reversed forks, reduced the number of insertions by more than half (Fig. 4d) and sensitized *dna2*Δ mutants to DNA damage (Extended Data Fig. 5). Yen1 nuclease cleaves 5′ flaps or stalled and reversed forks[28]. This nuclease is essential for growth in cells carrying hypomorphic mutants of Dna2[29], and constitutively active $yen1^{ON}$ could suppress lethality of *dna2*Δ[30] (Extended Data Fig. 5). This means that Yen1 can cleave at least some unprocessed structures in *dna2*Δ cells. Although *dna2*Δ *yen1*Δ cannot be constructed, we found that *dna2*Δ cells carrying constitutively active $yen1^{ON}$ exhibited increased numbers of insertions and complex events

© 2018 Springer Nature Limited. All rights reserved.

(Fig. 4d, Extended Data Fig. 5). These results suggest that Mus81 and Yen1 are needed for at least some of the insertions observed in Dna2-deficient cells. A model of insertions by alternative cleavage of structures normally processed by Dna2 is shown in Extended Data Fig. 6. To test whether low-molecular-weight DNA was present in Dna2-deficient cells, we separated DNA from wild-type and mutant cells by gel electrophoresis, extracted fragments corresponding to the size of insertions from the gel, and subjected them to quantitative PCR analysis using sets of primers specific for rDNA, the largest source of inserted DNA. We observed much higher levels of DNA in *pif1-m2 dna2*Δ cells compared to wild-type cells, and this difference was dependent on Fob1 (Extended Data Fig. 7).

The absence of Dna2 nuclease is not required for insertions, as long as extrachromosomal DNA is present in cells. Transformed double-stranded DNA or even single-stranded DNA was incorporated into DSBs in both wild-type and *pif1-m2 dna2*Δ cells (Extended Data Fig. 7). An increase of single-stranded DNA insertions in *pif1-m2 dna2*Δ cells suggests that Dna2 can also limit insertions by degrading single-stranded DNA formed in a Dna2-independent manner.

The mechanism of genome instability reported here may be related to the mechanism that causes insertions of comparable size in cancer[7], at V(D)J loci[10] and in the formation of short interstitial telomeric sequences common in the human genome, and may contribute to short gene duplications and chromosome evolution.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0769-8.

Received: 18 April 2018; Accepted: 23 October 2018;
Published online 5 December 2018.

1. Moore, J. K. & Haber, J. E. Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. *Nature* **383**, 644–646 (1996).
2. Teng, S. C., Kim, B. & Gabriel, A. Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. *Nature* **383**, 641–644 (1996).
3. Yu, X. & Gabriel, A. Patching broken chromosomes with extranuclear cellular DNA. *Mol. Cell* **4**, 873–881 (1999).
4. Morrish, T. A. et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.* **31**, 159–165 (2002).
5. Ricchetti, M., Fairhead, C. & Dujon, B. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402**, 96–100 (1999).
6. Onozawa, M. et al. Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. *Proc. Natl Acad. Sci. USA* **111**, 7729–7734 (2014).
7. Li, Y. et al. Patterns of structural variation in human cancer. Preprint at https://www.biorxiv.org/content/early/2017/08/27/181339 (2017).
8. Ju, Y. S. et al. Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. *Genome Res.* **25**, 814–824 (2015).
9. Henssen, A. G. et al. PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat. Genet.* **49**, 1005–1014 (2017).
10. Pieper, K. et al. Public antibodies to malaria antigens generated by two *LAIR1* insertion modalities. *Nature* **548**, 597–601 (2017).
11. Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **16**, 2164–2173 (1996).
12. Budd, M. E., Reis, C. C., Smith, S., Myung, K. & Campbell, J. L. Evidence suggesting that Pif1 helicase functions in DNA replication with the Dna2 helicase/nuclease and DNA polymerase δ. *Mol. Cell. Biol.* **26**, 2490–2500 (2006).
13. Belton, J. M. et al. The conformation of yeast chromosome III is mating type dependent and controlled by the recombination enhancer. *Cell Reports* **13**, 1855–1867 (2015).
14. Winston, F., Durbin, K. J. & Fink, G. R. The SPT3 gene is required for normal transcription of Ty elements in *S. cerevisiae*. *Cell* **39**, 675–682 (1984).
15. Sundararajan, A., Lee, B. S. & Garfinkel, D. J. The Rad27 (Fen-1) nuclease inhibits Ty1 mobility in *Saccharomyces cerevisiae*. *Genetics* **163**, 55–67 (2003).
16. Weitao, T., Budd, M., Hoopes, L. L. & Campbell, J. L. Dna2 helicase/nuclease causes replicative fork stalling and double-strand breaks in the ribosomal DNA of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **278**, 22513–22522 (2003).
17. Greenfeder, S. A. & Newlon, C. S. Replication forks pause at yeast centromeres. *Mol. Cell. Biol.* **12**, 4056–4066 (1992).
18. Makovets, S., Herskowitz, I. & Blackburn, E. H. Anatomy and dynamics of DNA replication fork movement in yeast telomeric regions. *Mol. Cell. Biol.* **24**, 4019–4031 (2004).
19. Gan, W. et al. R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev.* **25**, 2041–2056 (2011).
20. Markiewicz-Potoczny, M., Lisby, M. & Lydall, D. A critical role for Dna2 at unwound telomeres. *Genetics* **209**, 129–141 (2018).
21. Li, Z. et al. hDNA2 nuclease/helicase promotes centromeric DNA replication and genome stability. *EMBO J.* **209**, e96729 (2018).
22. Hu, J. et al. The intra-S phase checkpoint targets Dna2 to prevent stalled replication forks from reversing. *Cell* **149**, 1221–1232 (2012).
23. Thangavel, S. et al. DNA2 drives processing and restart of reversed replication forks in human cells. *J. Cell Biol.* **208**, 545–562 (2015).
24. Liu, B., Hu, J., Wang, J. & Kong, D. Direct visualization of RNA-DNA primer removal from Okazaki fragments provides support for flap cleavage and exonucleolytic pathways in eukaryotic cells. *J. Biol. Chem.* **292**, 4777–4788 (2017).
25. Pike, J. E., Burgers, P. M., Campbell, J. L. & Bambara, R. A. Pif1 helicase lengthens some Okazaki fragment flaps necessitating Dna2 nuclease/helicase action in the two-nuclease processing pathway. *J. Biol. Chem.* **284**, 25170–25180 (2009).
26. Stith, C. M., Sterling, J., Resnick, M. A., Gordenin, D. A. & Burgers, P. M. Flexibility of eukaryotic Okazaki fragment maturation through regulated strand displacement synthesis. *J. Biol. Chem.* **283**, 34129–34140 (2008).
27. Lee, M. et al. Rad52/Rad59-dependent recombination as a means to rectify faulty Okazaki fragment processing. *J. Biol. Chem.* **289**, 15064–15079 (2014).
28. Blanco, M. G., Matos, J. & West, S. C. Dual control of Yen1 nuclease activity and cellular localization by Cdk and Cdc14 prevents genome instability. *Mol. Cell* **54**, 94–106 (2014).
29. Olmezer, G. et al. Replication intermediates that escape Dna2 activity are processed by Holliday junction resolvase Yen1. *Nat. Commun.* **7**, 13157 (2016).
30. Michel, A. H. et al. Functional mapping of yeast genomes by saturated transposition. *eLife* **6**, e23570 (2017).

Acknowledgements We thank A. Gabriel, D. J. Garfinkel, J. Haber, M. G. Blanco and F. Storici for the gifts of strains and plasmids, and J. Haber and P. Hastings for critical reading of the manuscript. This work was funded by grants from the US National Institutes of Health (GM080600 and GM125650 to G.I., GM125632 and HL133254 to K.C.) and the Cancer Prevention Research Institute of Texas (RP140456 to G.I. and G.P., RP150611 to K.C.).

Reviewer information *Nature* thanks P. Cejka, L. Symington and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author Contributions Y.Y., N.P. and B.X. contributed equally to this work. Y.Y., N.P. and Z.Y. constructed strains; Y.Y. and A.P. performed all experiments related to insertions at DNA breaks; N.P. carried out experiments on transposition; B.X., G.W. and K.C. designed, performed and described bioinformatics analysis. G.I., Y.Y. and G.P. designed the experiments, discussed the data and wrote the manuscript.

Competing Interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0769-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0769-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.C. or G.I.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Media, yeast strains and plasmids. All strains used in this work are derivatives of three strains: (i) JKM139 to study insertion at *MAT**a*** locus (DELho *hml*::*ADE1 MAT**a*** *hmr*::*ADE1 ade1 leu2-3,112 lys5 trp1*::*hisG ura3-52 ade3*::*GAL10*::*HO*)[11]; (ii) yYY379 to study insertion at *URA3* locus (DELho *hml*::*ADE1 MAT**a***::*hphMX hmr*::*ADE1 ade1 leu2-3,112 lys5 trp1*::*hisG URA3*::*actin intron*::*HOcs ade3*::*GAL10*::*HO*); and (iii) DG1657 to study retrotransposition and Ty1 cDNA level (*MAT**a*** *ura3-167 his3Δ-200 trp1-hisG leu2-hisG Ty1-270his3-AI Ty1-588neo Ty1-146[tyb1*::*lacZJ]*)[15], a gift from D. J. Garfinkel. yYY379 strain was obtained by replacing the HO cleavage site with *hphMX* at *MAT**a*** locus and by replacing *URA3* with *URA3*::*ACT1 intron*::*HO*cs. *URA3*::*ACT1 intron*::*HO*cs cassette was amplified from AGY117 strain[3], a gift from A. Gabriel. A list of all strains is presented in Supplementary Information Table 2. Helicase-dead mutant (*dna2*^{R1253Q}) and nuclease-dead mutant (*dna2*^{E675A}) of *DNA2* were introduced into the genome using the *delitto perfetto* approach[31].

HO induction and analysis of NHEJ efficiency. To induce HO endonuclease, cells from an overnight saturated culture in YEPD (yeast extract–peptone–dextrose) (1% yeast extract, 2% peptone, 2% dextrose) were washed twice with YEP–raffinose (1% yeast extract, 2% peptone, 2% raffinose), inoculated into 5 ml YEP–raffinose and incubated overnight at 30 °C. When the density of the culture reached ~1–2 × 10^7 cells/ml, cells were spread on YEP (yeast extract, peptone)–galactose plates (1% yeast extract, 2% peptone, 2% galactose) and incubated at 30 °C for up to 6 days. As a control, cells were spread onto YEPD plates. The NHEJ efficiency was calculated as the number of colonies on YEP–galactose divided by the number of colonies on YEPD. The experiment was repeated at least three times for each mutant. For hydroxyurea treatment, hydroxyurea was added to a final concentration of 80 mM when the density of the culture reached ~1 × 10^7 cells/ml in YEP-raffinose and incubated for 4 h before plating.

Analysis of insertions at *MATa* locus. Single colonies from YEP–galactose plates were used for colony PCR using the following primers: mata-F (ACTTCAAGTAAGAGTTTGGGTATGT) (165 bp upstream of HO cleavage site) and mat-Rw (TACTGACAACATTCAGTACTCGAAAG) (165 bp downstream of HO cleavage site). The amfiSure PCR Master Mix (GenDEPOT, cat. no. P0311) was used for PCR with the following conditions: 94 °C for 5 min; 35 cycles of 94 °C for 30 s, 52 °C for 30 s and 72 °C for 2 min 30 s. PCR products were analysed by electrophoresis (1.2% agarose in 1× TBE buffer) at 8 V/cm for 30 min. PCR products having large insertions were cleaned up with the NucleoSpin Kit (Macherey-Nagel, cat. no. 740609) and sequenced by Sanger sequencing. ApE software was used to analyse the microhomology of insertion. SnapGene was used to map the insertion to chromosome, Ty, rDNA and 2μ plasmid. To determine statistically significant differences in insertion frequencies between strains we used χ^2 analysis.

Analysis of insertion at *URA3* locus. Yeast were grown in YEP–raffinose up to a density of ~1–2 × 10^7 cells/ml, and galactose was added to a final concentration of 2% and incubated at 30 °C for 24 h. Cells were plated on 5-fluoroorotic acid (5-FOA) plates and incubated at 30 °C for 6 days. For transient DSB induction glucose was added at 1, 2 or 4 h to a final concentration of 2% to shut down the expression of galactose-inducible HO and the cells were plated on 5-FOA plates and incubated at 30 °C for 6 days. To screen for insertions at the DSB, primers Act1-Fw (ATATCGTGGTTATTACAGATCAGTCA) (165 bp upstream of HO cleavage site) and Ura3-Rw (ATTGTTAGCGGTTTGAAGCAGG) (165 bp downstream of HO cleavage site) were used. The sequencing and analysis of inserts was performed as described above for the *MAT**a*** locus.

Analysis of insertion at *LYS2* locus. Plasmids marked with the *LEU2* gene containing constitutively expressed gRNA gLYS2-2^W and a galactose-inducible Cas9, a gift from J. Haber[32], were transformed into wild-type and mutant yeast cells. To induce Cas9, cells from an overnight saturated culture in leucine drop-out glucose medium were washed and inoculated into 5 ml YEP–raffinose and incubated overnight at 30°C. When the density of the culture reached ~1–2 × 10^7 cells/ml, cells were spread on leucine drop-out galactose plates (2% galactose) and incubated at 30 °C for up to 6 days. As a control, cells were spread onto leucine drop-out-glucose plates. The NHEJ efficiency was calculated as described[32]. To test for large insertions, single colonies from leucine drop-out galactose plates were used for colony PCR using the following primers: Lys2-Fw (TAGACGAGTTCAAGCATCATTTAGT) (120 bp upstream of Cas9 cleavage site) and Lys2-Rw (CAAGTTCTTAGTTGGATCAGGT) (122 bp downstream of Cas9 cleavage site). PCR fragments carrying insertions were sequenced and analysed.

Analysis of extrachromosomal DNA. Yeast were grown in YEPD to a density of ~1–2 × 10^7 cells/ml. Then, 1 × 10^8 cells were collected and washed twice in buffer A (100 mM EDTA, 50 mM Tris-HCl, pH 7.4). Cells were resuspended in melted (45 °C) 0.5% agarose prepared in buffer A with 0.3 mg/ml Zymolyase 100T (in 10 mM KPO_4, pH 7.4) and transferred to a plug mould. Plugs were incubated in 400 μl buffer A at 37 °C for 1 h followed by addition of 100 μl 5% sarcosyl for 1.5 h and followed by addition of 10 μl 10 mg/ml RNase A and 4 μl Riboshredder (Epicentre, cat. no. RS12500) for an additional 2.5 h. The plugs containing DNA were loaded onto 0.8% agarose and electrophoresis was performed in 1× TBE buffer at 8 V/cm for 40 min. In each lane the agarose gel fragments corresponding to 75 bp–1.5 kb and to slow migrating genomic DNA were cut separately. The DNA was extracted using the NucleoSpin Kit. The Power SYBR Green PCR Master Mix (Applied Biosystems, cat. no. 4367659) was used for quantitative PCR with the following conditions: 95 °C for 10 min; 40 cycles of 95 °C for 15 s, 60 °C for 1 min. Primers were NTS1-QF2 (TGGCTTCCTATGCTAAATCCC) and NTS1-QR3 (GCATAATGGAGTGCTTAACTCTTC) for NTS1#A, NTS1-QF (ACACCCTCGTTTAGTTGCTTC) and NTS1-QR (CGGTATGCGGAGTTG TAAGATG) for NTS1#B. The amount of short DNA (75 bp–1.5 kb) from each lane was normalized to large mass genomic DNA. The cycle threshold numbers were used to determine the fold difference in the amount of short DNA between wild type and mutants.

Analysis of insertion of transformed DNA. To analyse the efficiency of insertions of transformed DNA into a DSB the 98-nt-long oligonucleotides M13-X-98-Fw (GTACTAAGACTCATAATTACATTTGGCGTTATGTATCTGCATTAGTTGA ATGTGGTATTCCTAAATCTCAACTGATGAAACGTTCAACGTGACAAGTC) and complementary M13-X-98-Rw (GACTTGTCACGTTGAACGTTT CATCAGTTGAGATTTAGGAATACCACATTCAACTAATGCAGATACATAA CGCCAAATGTAATTATGAGTCTTAGTAC) were synthesized. To obtain duplex DNA, equal amounts of complementary oligonucleotides were mixed and heated at 95 °C for 5 min and cooled down at room temperature. Reverse-phase cartridge-purified oligonucleotides were purchased from Sigma-Aldrich and dissolved with annealing buffer (1 mM EDTA, 50 mM NaCl, 10 mM Tris, pH 7.5) to 200 μM. When culture density reached ~1–2 × 10^7 cells/ml in YEP–raffinose, cells were collected and washed twice with water. For each transformation, 1.5 × 10^8 cells were mixed with 240 μl 50% PEG3350, 36 μl 1M lithium acetate and 20 μl 100 μM double-stranded DNA (dsDNA) or 40 μl 100 μM single-stranded DNA (ssDNA) and water to total 360 μl. The transformation mixture was incubated at 30°C for 30 min followed by 42 °C for 30 min. The cells were centrifuged, resuspended in water, spread on galactose plates and incubated at 30 °C for 6 days. The primers mata-F and mat-Rw were used to test insertion of transformed DNA. All insertions were sequenced and analysed.

Analysis of Ty cDNA amount. A single colony from each mutant was inoculated into 5 ml of YEPD and incubated overnight at 24 °C. A 5-μl aliquot of each culture was inoculated into 25 ml of fresh YEPD the next day and grown for an additional 2 days at 24 °C. Cells were collected and total genomic DNA was isolated by glass bead disruption using a standard phenol chloroform extraction protocol. DNA was then digested with PvuII and separated on 0.8% agarose gels. Southern blotting and hybridization with radiolabelled DNA probes was carried out using a standard protocol[33]. Ty1 PvuII-SnaBI fragment of Ty1-H3 was used as a ^{32}P-labelled DNA probe and was prepared by PCR using primers *PvuII*-F1 (CTGTAAAAGCAGTAAAATCAATCAAAC) and *Sna*BI-R1 (GTATAGA TTATTACCTGATACTTCATCTCT). Intensity of bands on Southern blots corresponding to probed DNA fragments was analysed using ImageQuant TL software (Amersham Biosciences) and normalized to three genomic Ty1 fragments, as previously described[15].

Analysis of Ty cDNA stability. A single colony from each mutant was inoculated in 80 ml of YEPD liquid medium for 2 days at 24 °C. Twenty millilitres of this culture was then pelleted, washed, and resuspended in 250 ml of fresh YEPD. Cultures were then shaken at 24 °C for 2 h. Aliquots (50 ml) were removed (time 0) before adding the reverse transcriptase inhibitor phosphonoformic acid (PFA; Sigma) to a concentration of 600 μg/ml as previously described[34]. Aliquots (50 ml) aliquots were removed after 1, 2, 4 or 6 h of growth at 24 °C. Cells were then pelleted, and total genomic DNA was isolated, digested with PvuII, and processed for Southern analysis as described above. The stability of cDNA was calculated from the slope of the best-fit line. The amount of DNA in each lane was normalized using bands corresponding to genomic Ty. The amount of cDNA over time was quantified by dividing pixel intensities of the bands corresponding to cDNA with PFA by the intensity of the bands corresponding to cDNA without PFA in each time point. All quantification was done using ImageQuant TL software.

Analysis of Ty1 transcription. A single colony from each mutant was inoculated in 5 ml of YEPD overnight at 24 °C. A 5-μl aliquot of each culture was inoculated into 25 ml of fresh YEPD and grown for an additional 2 days at 24 °C. A 5-ml aliquot of each culture was pelleted, and RNA was extracted using the MasterPure Yeast RNA Purification Kit (Lucigen cat. no. MPY03100). RNA samples mixed with glyoxal loading dye were separated on a 1% agarose gel and transferred to nitrocellulose membrane using the NorthernMax-Gly Kit (Invitrogen cat. no. AM1946). The ^{32}P-labelled DNA probes were made by randomly primed DNA synthesis. Ty1 PvuII-SnaBI fragment of Ty1-H3 was used as a ^{32}P-labelled DNA probe and was prepared as described above. The control PYK1 probe was prepared by PCR using

© 2018 Springer Nature Limited. All rights reserved.

two primers PYK1-F1 (GTTGTTGCTGGTTCTGACTTGAGAA) and PYK1-R1 (TCAAGATACCGAATTCCTTAGCC). The intensity of bands on Southern blots corresponding to Ty RNA fragments was analysed with ImageQuant TL and normalized to the PYK1 RNA signal.

Analysis of Ty retrotransposition rates. The rate of retrotransposition was estimated in strains carrying the Ty1-270*his3-AI* reporter[15]. Wild-type and mutant cells were streaked for single colonies on YEPD plates. Individual colonies were used to inoculate 5 ml YEPD cultures that were incubated at 24 °C, diluted and grown to 1×10^7 cells/ml before plating 20–30 cells per plate on YEPD. Plates were incubated at 24 °C for 6 days. For each strain tested, 10–20 individual colonies were diluted in water and spread on synthetic complete medium lacking histidine and incubated at 24 °C for 3–4 days. His^+ colonies were then counted. To perform statistical comparisons of spontaneous transposition rates between genotypes we used the Drake estimator as previously described[35]. The bootstrap resampling approach was used to determine *P* values.

Bioinformatic analysis of genomic features related to insertion sites. Positions of confirmed origins of replication (ARSes) were downloaded from the OriDB database (http://cerevisiae.oridb.org/)[36]. R-loop reference positions were collected from a published source[37]. Hi-C interaction maps were collected from the GEO database (with accession numbers GSM1905067 and GSM1905068)[13]. All other genome features were acquired from the SGD database (https://downloads.yeastgenome.org/). Random control insertions were created based on the size and distribution of real insertions. Specifically, for each real insertion a corresponding control insertion of the same size with a random location and on the same chromosome was generated. For analysis of hotspots (multiple insertion donor DNAs within a region of 3 kb or shorter), a random hotspot with the same size, distance and on the same chromosome was generated. The randomization was repeated 1,000 times and the mean value for each type of analysis was taken for comparison with observed inserted DNA. All values for random control insertions are the average of 1,000 different sets of random insertion controls. Overlapped events between insertions and genome features are defined as having at least 1-bp overlap. For distance analysis, the distance of each insertion is calculated by the edge distance between the insertion and its closest genome feature. If an insertion is overlapped with a genomic feature, then its distance is defined as 1 bp. For analysis presented in Fig. 4a–c, all insertions coming from Ty retrotransposons, rDNA or 2μ plasmid are excluded. *P* values in bar graphs were calculated by an empirical cumulative distribution function based on 1,000 times bootstrapping (Fig. 4a–c). *P* values in box plots were determined based on one-tailed Wilcoxon test (Fig. 4a–c). Correlation coefficients were calculated based on the Spearman method. For bar graphs (Fig. 4a, b), error bars represent the standard deviation. For box plots presented in Fig. 4a–c, e, the centre line represents the median value, the bottom and top of the box are upper and lower quartiles, the upper whisker is located at the minimum value of maximum value and (Q3 + 1.5 × interquartile range) (IQR), and the lower whisker is located at the maximum of minimum value and (Q1 − 1.5 × IQR) (Q1 is first quartile, Q3 is third quartile, IQR is interquartile range). For Hi-C interaction map analysis (Extended Data Fig. 2), insertions of Ty retrotransposons, telomeres, rDNA and 2μ plasmid are excluded from analysis. A two-tailed Wilcoxon test was used to determine whether the frequency of interaction between the HO cleavage site and donor DNA sites is significantly different when compared to the interaction between the HO cleavage site and randomly selected loci. The randomization was repeated 1,000 times and the median *P* value was used to determine the significance of the difference.

Code availability. All codes used in this project are deposited at https://github.com/fagisX/FAID.

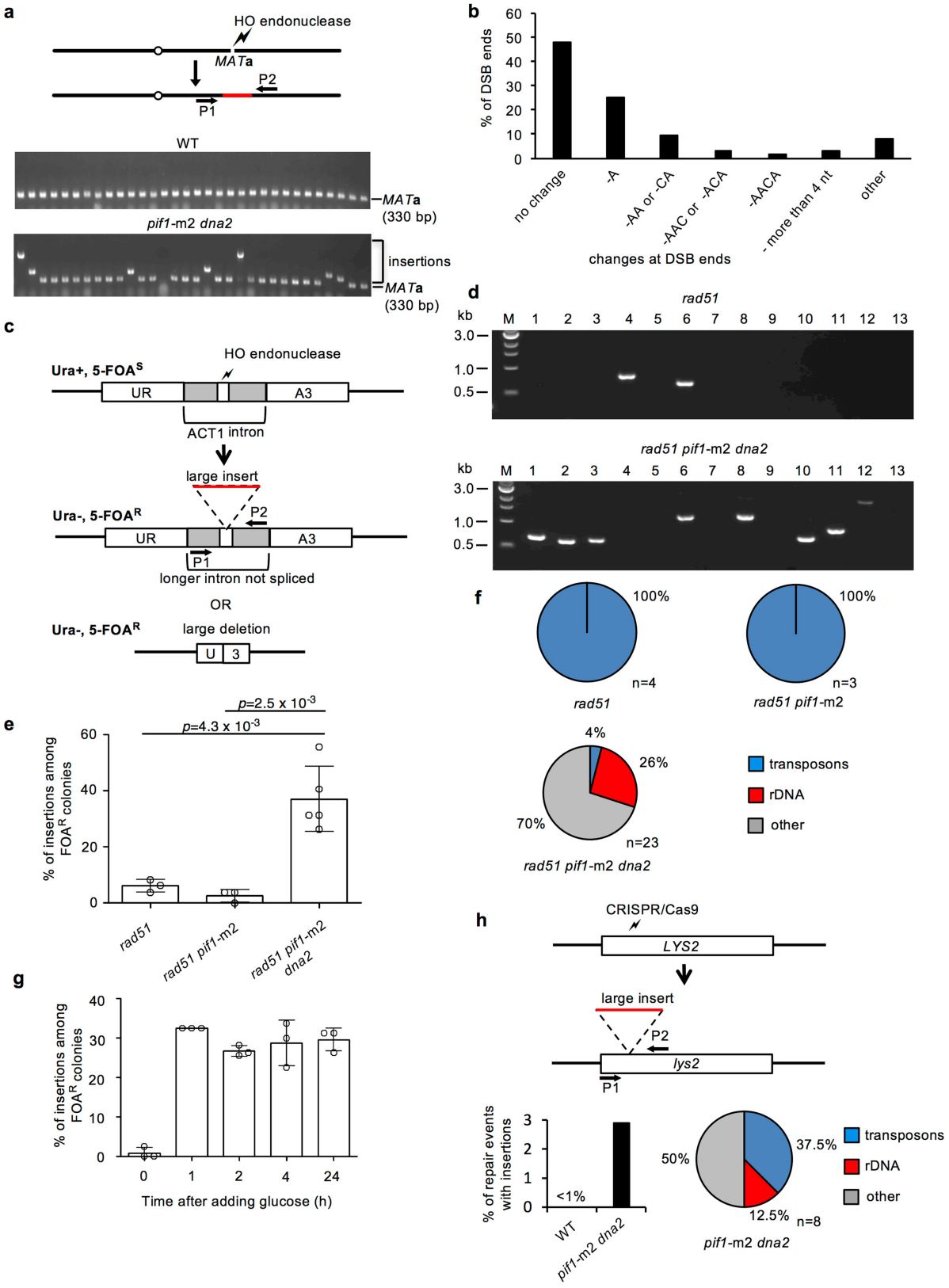
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data supporting the findings of this study are available within the Letter. Sequences of all inserted DNA and sequences of the junctions analysed are provided in Supplementary Table 1. Source gel images are presented in Supplementary Figure 1.

31. Storici, F. & Resnick, M. A. The *delitto perfetto* approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. *Methods Enzymol.* **409**, 329–345 (2006).
32. Lemos, B. R. et al. CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proc. Natl Acad. Sci. USA* **115**, E2040–E2047 (2018).
33. Church, G. M. & Gilbert, W. Genomic sequencing. *Proc. Natl Acad. Sci. USA* **81**, 1991–1995 (1984).
34. Lee, B. S., Bi, L., Garfinkel, D. J. & Bailis, A. M. Nucleotide excision repair/TFIIH helicases RAD3 and SSL2 inhibit short-sequence recombination and Ty1 retrotransposition by similar mechanisms. *Mol. Cell. Biol.* **20**, 2436–2445 (2000).
35. Mayle, R. et al. DNA repair. Mus81 and converging forks limit the mutagenicity of replication fork breakage. *Science* **349**, 742–747 (2015).
36. Siow, C. C., Nieduszynska, S. R., Muller, C. A. & Nieduszynski, C. A. OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res.* **40**, D682–D686 (2012).
37. Wahba, L., Costantino, L., Tan, F. J., Zimmer, A. & Koshland, D. S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev.* **30**, 1327–1338 (2016).

© 2018 Springer Nature Limited. All rights reserved.

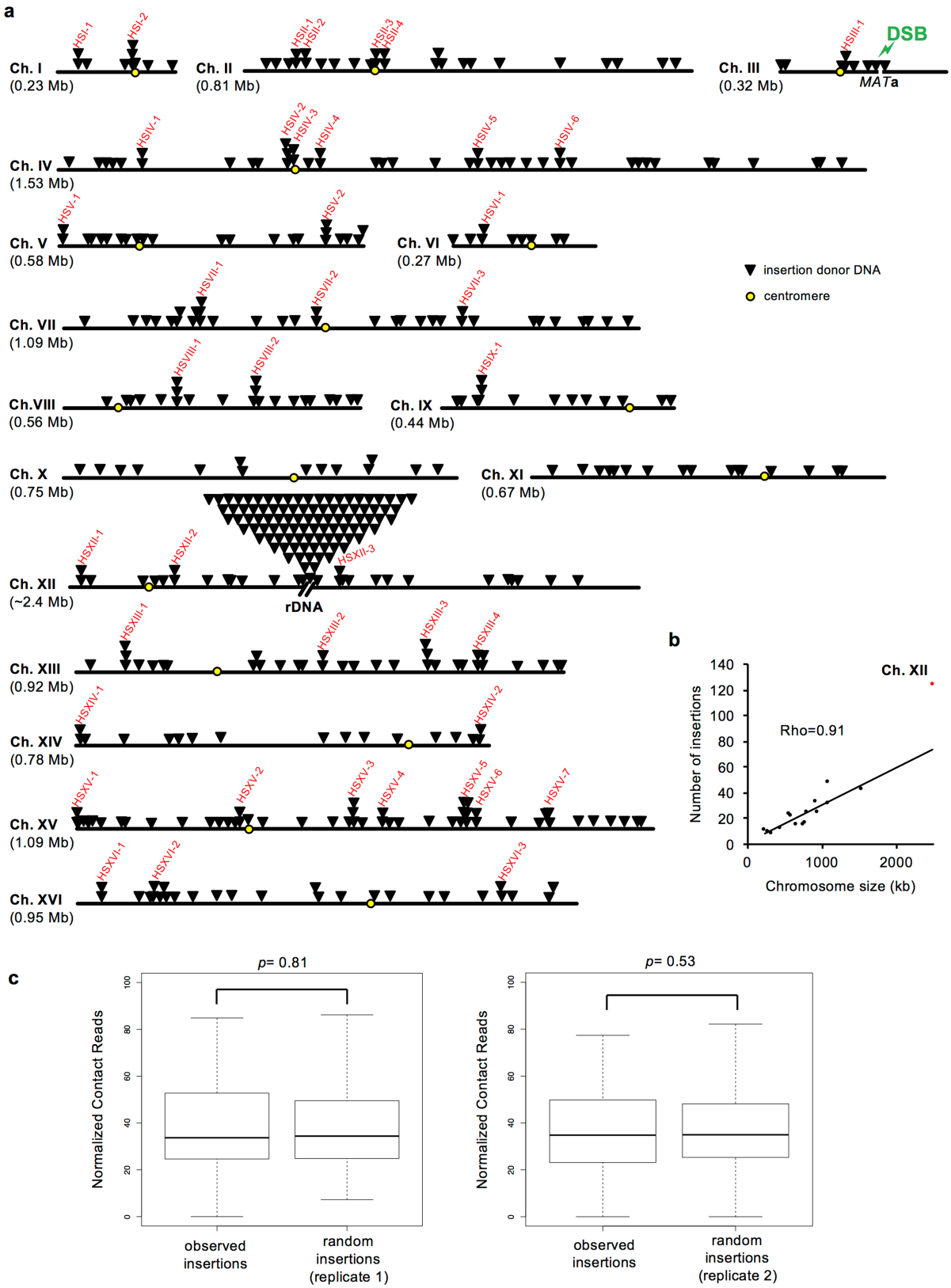


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

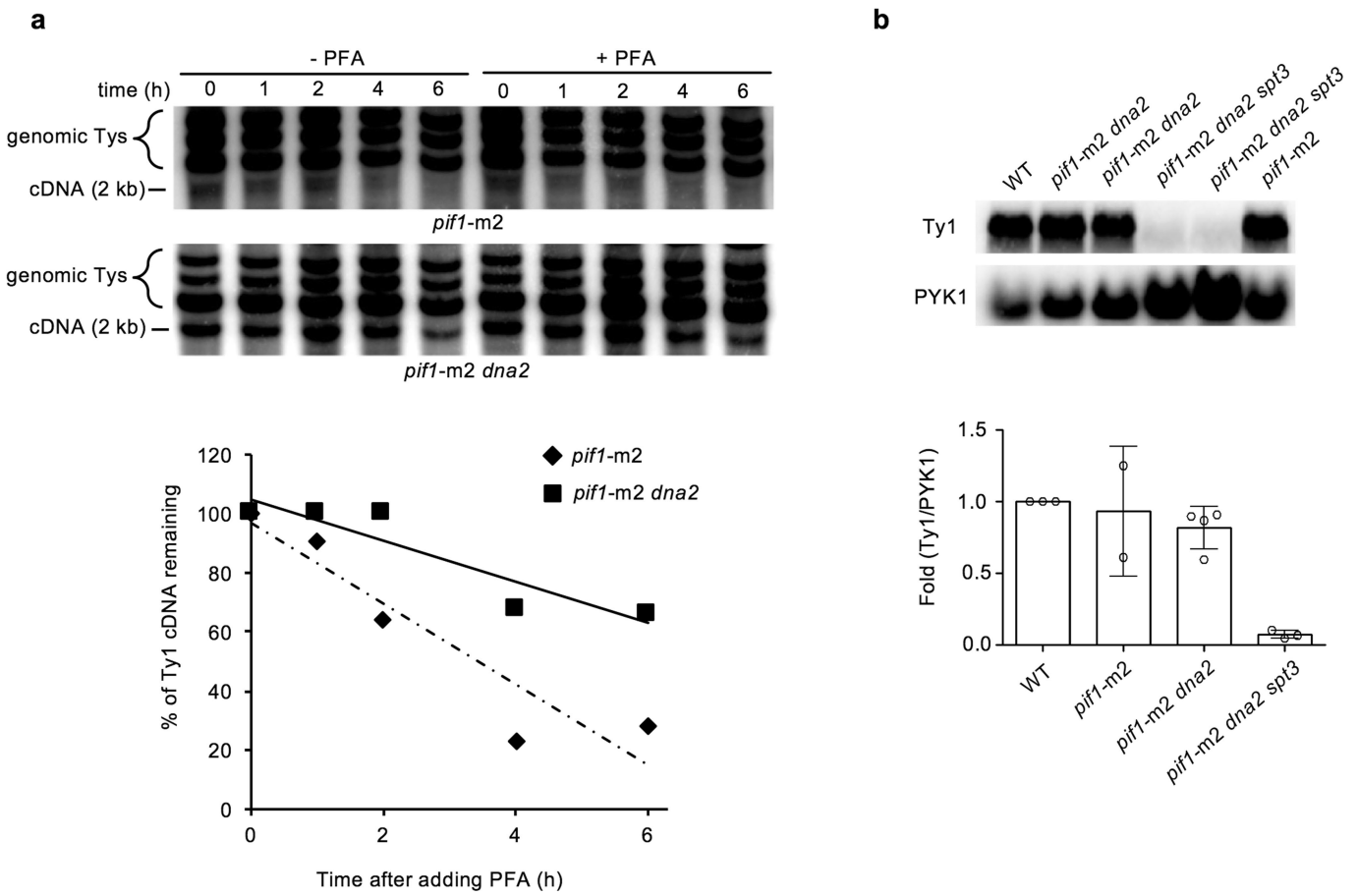
Extended Data Fig. 1 | Insertion analysis at *MATa*, *URA3* and *LYS2* loci. **a**, Experimental system to study insertions at DSBs and PCR analysis of *MAT**a*** locus after DSB repair in wild-type and *pif1-m2 dna2* cells. Analysis was repeated more than 10 times (for gel source data, see Supplementary Figure 1). **b**, Analysis of change of DSB ends among insertion events. **c**, Schematic showing experimental system. A HO break is generated at an *ACT1* intron integrated in the *URA3* gene. Insertion of a DNA fragment or large deletion interferes with splicing and generates uracil auxotrophs. **d**, Analysis of insertions by PCR and agarose gel electrophoresis at *URA3*. The experiment was repeated more than three times with similar results. For gel source data, see Supplementary Figure 1. **e**, Percentage of insertions among 5-FOA resistant colonies. Data are mean ± s.d.; $n = 3$ independent experiments; two-tailed *t*-test. **f**, Analysis of origin of DNA inserted at DSB at *URA3* locus in indicated mutants, *n* represents number of independent insertions from indicated mutants. **g**, Percentage of insertions among 5-FOA resistant colonies after transient induction of HO break in *rad51 pif1-m2 dna2*. Data are mean ± s.d.; $n = 3$ independent experiments. **h**, Schematic showing experimental system to follow insertions at CRISPR–Cas9-induced DSBs within the *LYS2* locus. Below, percentage of insertions among cells maintaining CRISPR–Cas9 and analysis of origin of DNA inserted at *LYS2*. *n* represents number of independent insertions sequenced in *pif1-m2 dna2* cells. The experiment was repeated four times with similar results.

© 2018 Springer Nature Limited. All rights reserved.



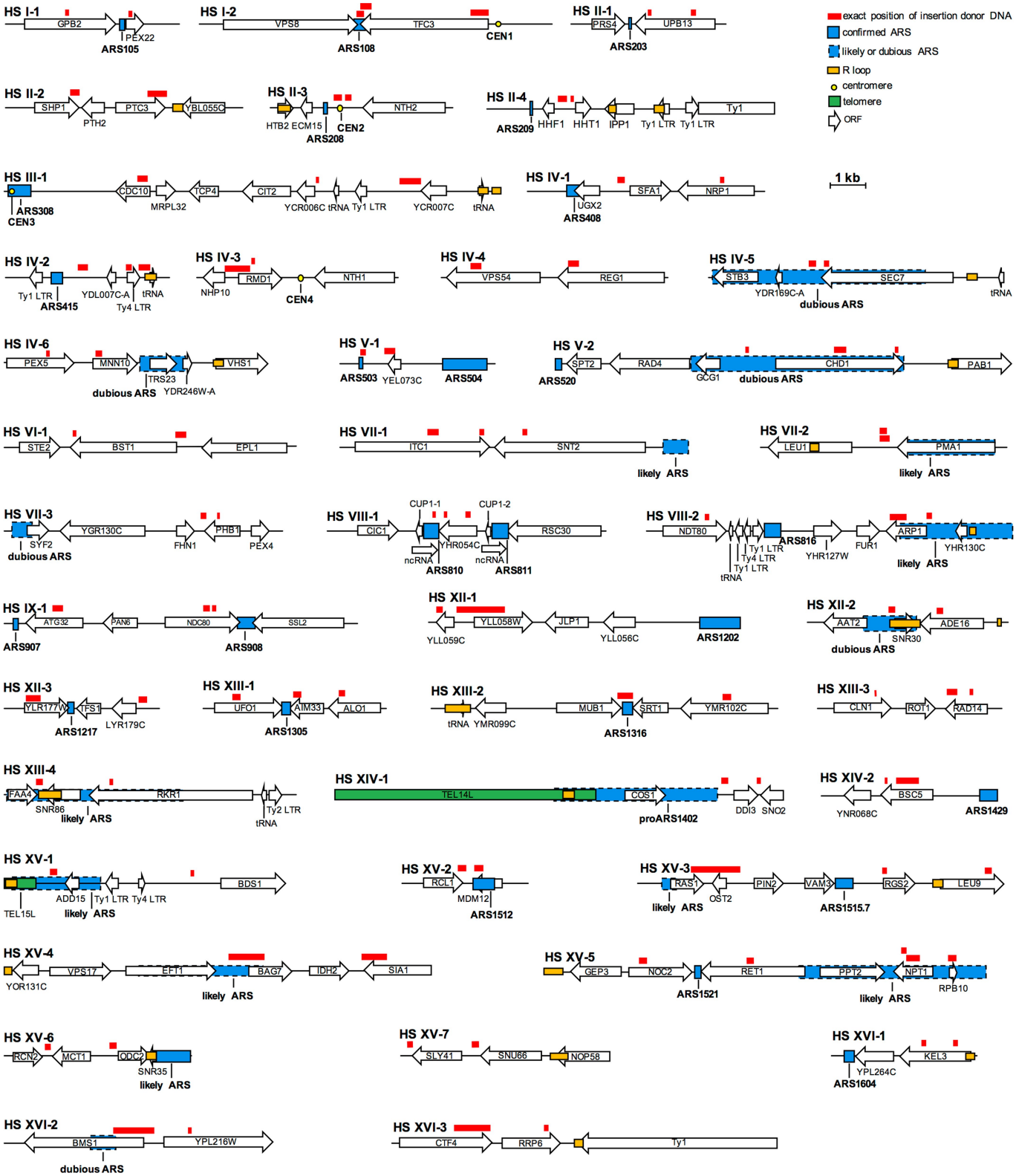
Extended Data Fig. 2 | Origin of inserted DNA at DSBs. a, Each triangle indicates a single insertion donor DNA; hotspots of insertion donor DNA are marked in red. **b**, Scatter plot of chromosome size and insertion number. $n = 468$ independent insertions. Correlation coefficients were calculated based on the Spearman method. **c**, Contact analysis between *MAT* locus on chromosome III and loci from which DNA was inserted. For each replicate, 1,000 random sets of DNAs (equal size and number) are compared to experimental inserted DNA. *P* values are determined by two-tailed Wilcoxon test; $n = 358$ independent inserted DNAs used for contact analysis.

© 2018 Springer Nature Limited. All rights reserved.



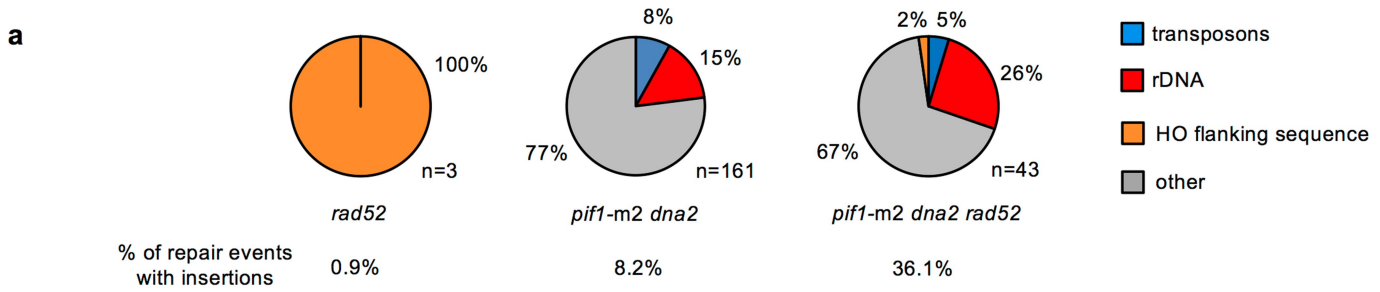
Extended Data Fig. 3 | Analysis of transposon cDNA stability and Ty1 expression. a, Analysis of Ty1 cDNA stability. The experiment was repeated four times with similar results. **b**, Analysis of Ty1 expression and its quantification. Data are mean ± s.d. from three independent experiments. For gel source data, see Supplementary Figure 1.

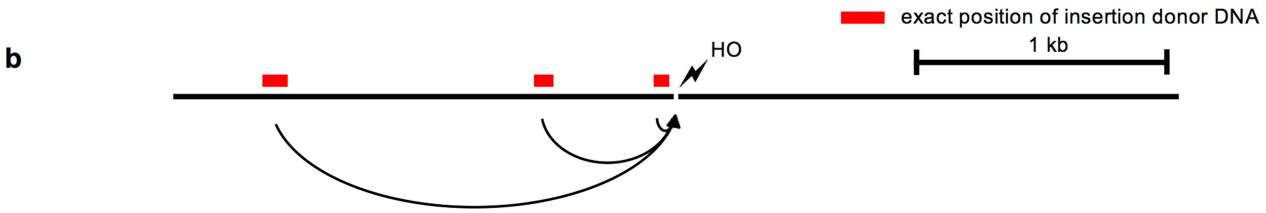
© 2018 Springer Nature Limited. All rights reserved.

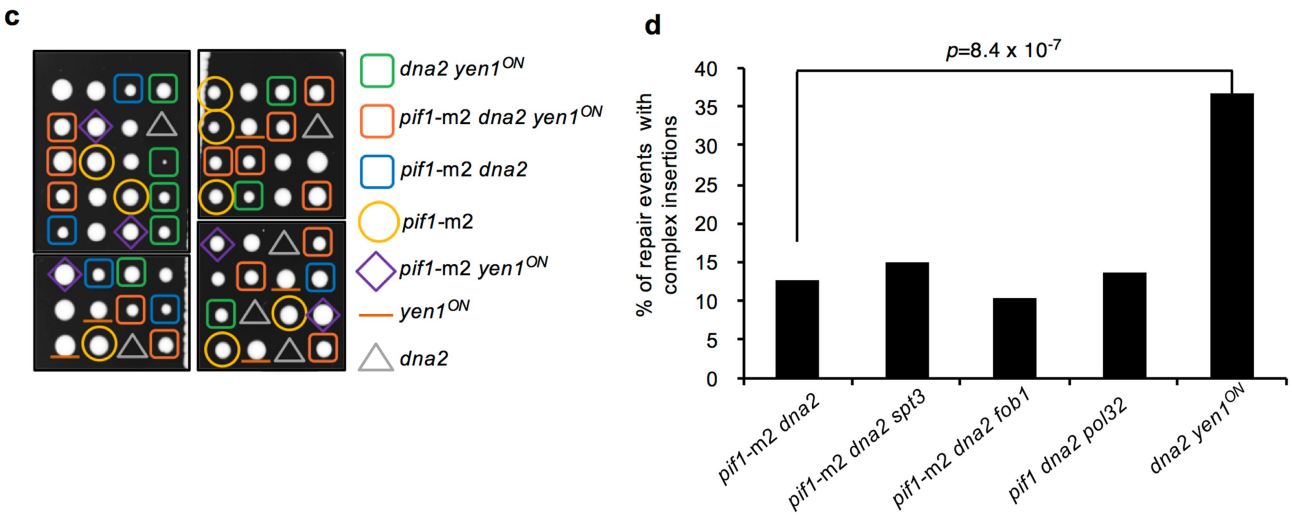


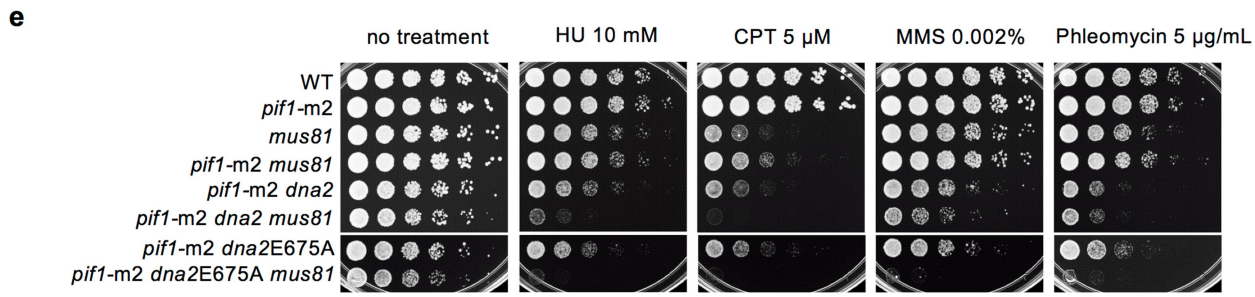
Extended Data Fig. 4 | Hotspots of origin of inserted DNAs. Position of DNAs inserted within DSBs is indicated in red. Blue boxes, origins of replication; yellow circles, centromeres; green boxes, telomeres; open boxes, genes. Hotspots were defined as loci that are the source of at least two inserted DNA fragments separated from each other by no more than 3 kb.

© 2018 Springer Nature Limited. All rights reserved.



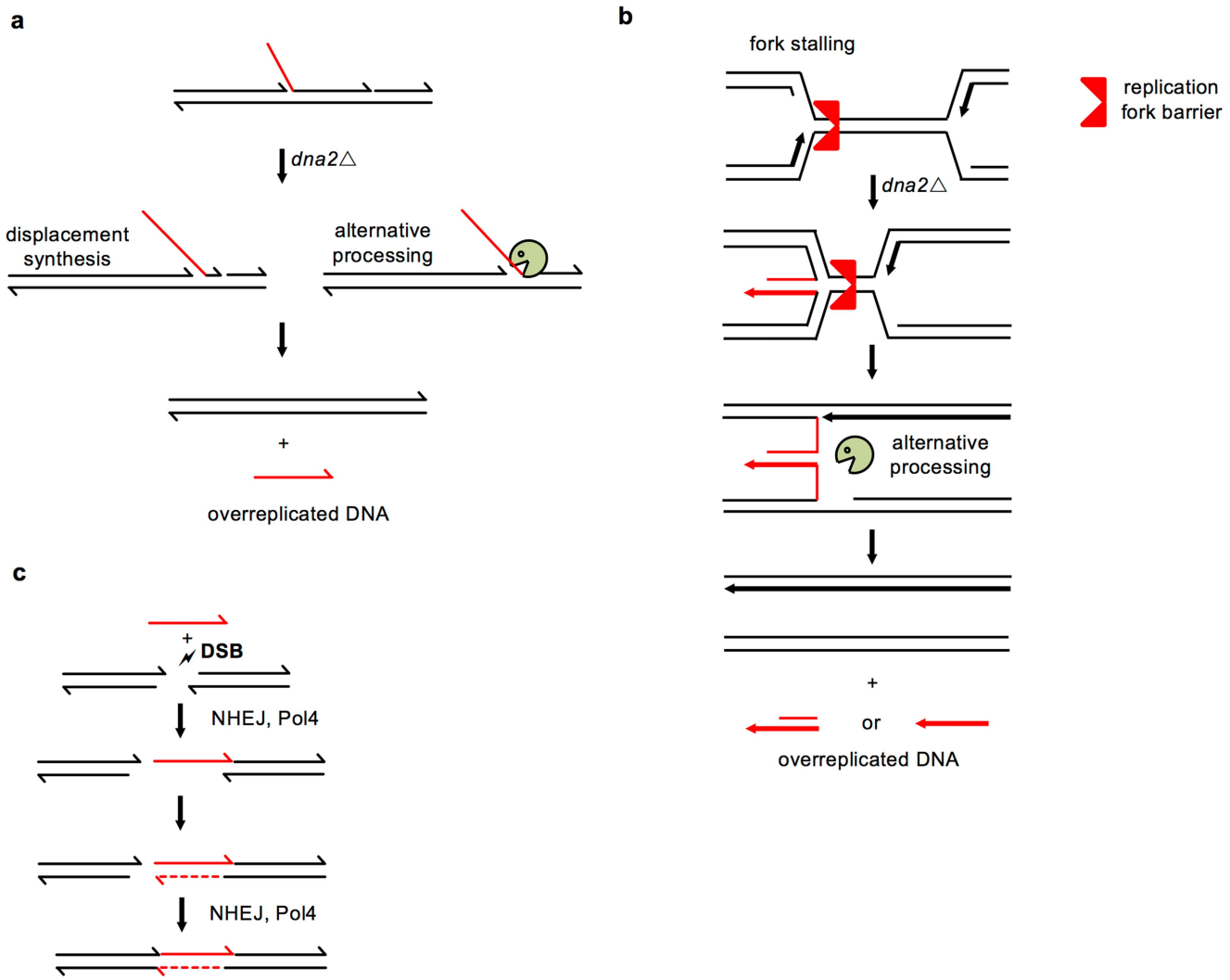






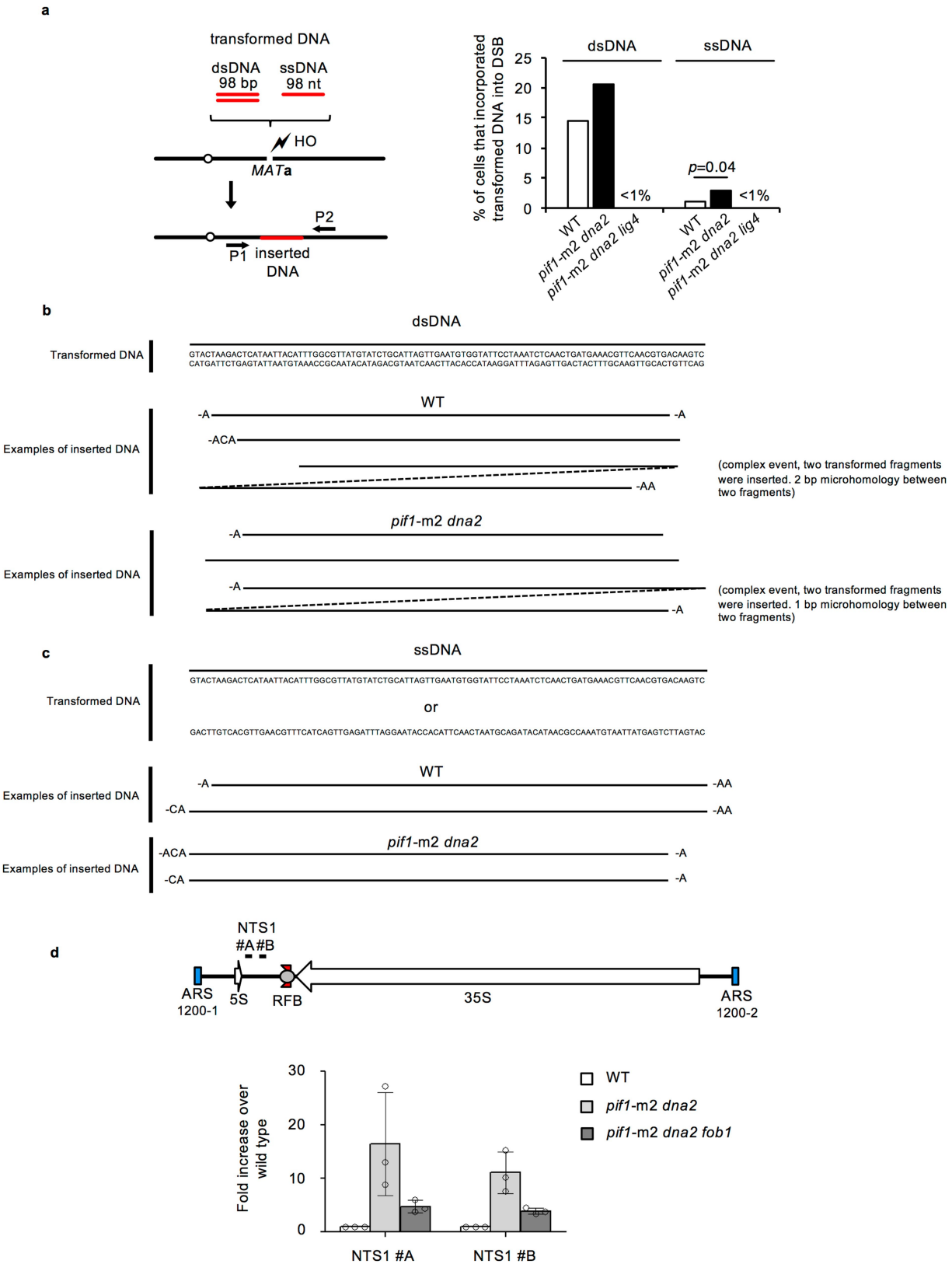
Extended Data Fig. 5 | Genetic interactions between Dna2 and Rad52, Yen1 and Mus81. a, Overall frequency and analysis of origin of DNA inserted at DSB in indicated mutants. *n* represents the number of independent insertions analysed by sequencing. **b**, Origin of insertions in *rad52*Δ mutant cells. **c**, Active Yen1 rescues non-viability of *dna2*Δ cells. Tetrad dissection of *PIF1*/*pif1-m2 YEN1*/*yen1*^{ON} *DNA2*/*dna2*Δ triple heterozygotes is shown. The experiment was repeated twice. **d**, Analysis of complex insertions (2 or more DNA fragments inserted at DSB) in Dna2-deficient mutants. Sample size, defined as the number of independent insertions analysed for each mutant is presented in Extended Data Table 1. χ^2 test is used to determine the *P* value. **e**, DNA damage sensitivity analysis (spot assay, 5× dilution) in indicated mutants. The experiment was repeated twice.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Model of large insertions at DSBs in Dna2-deficient cells. a, Unprocessed 5′ flaps are processed by alternative nuclease or displaced by synthesis leading to release of over-replicated DNA fragments. **b**, Stalled and reversed forks, when approached by a converging fork, leave over-replicated DNA that can be released by processing by other nucleases. **c**, ssDNA can be inserted into DSBs by NHEJ and Pol4.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Analysis of insertions of transformed DNA at DSBs and analysis of free, short DNA in cells. a, Analysis of insertions of transformed DNA at DSBs in wild-type and indicated mutant cells. Schematic of the experiment (left) and percentage of cells carrying insertion (right). χ^2 test was used to determine the *P* values; $n = 160$ for dsDNA and $n = 320$ for ssDNA, and represents the number of colonies tested for the presence of insertion. **b**, **c**, Analysis of inserted DNA after transformation of dsDNA (**b**) and ssDNA (**c**). **d**, Quantitative PCR analysis of short free DNA in indicated mutants. Data are mean ± s.d.; $n = 3$ independent experiments. Position of the primers used is shown at the top and fold change in DNA amount is shown on the bottom.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Analysis of NHEJ efficiency and insertion frequency

Table 1a

Genotype	Number of insertions at *MAT***a** (%)*	NHEJ efficiency**	# of sequenced events	# of inserts ***	# of complex events***	transposon insertions (%) ***	rDNA fragment insertions (%) ***	other nuclear genome insertions (%)***
WT	0% (0/644)	0.104%±0.058%	N/A					
pif1-m2	0% (0/160)	0.061%±0.013%	N/A					
pif1-m2 *dna2*	8.2% (148/1794)	0.105%±0.029%	142	161	18	8.1%	14.9%	77.0%
pif1-m2 *dna2*R1253Q	0% (0/160)	0.116%±0.099%	N/A					
pif1-m2 *dna2*E675A	15.6% (50/320)	0.123%±0.059%	3					
rad51 pif1-m2 *dna2*	9.4% (30/320)	0.096%±0.020%	13					
rad51	0% (0/320)	0.079%±0.001%	N/A					
rad52 pif1-m2 *dna2*	36.1% (52/144)	0.228%±0.091%	33	43	8	4.7%	25.6%	69.7%
rad52	0.9% (3/320)	0.167%±0.037%	3					
rad27	0% (0/320)	0.193%±0.165%	N/A					
yku70 pif1-m2 *dna2*	0% (0/160)	0.011%±0.009%	N/A					
lig4 pif1-m2 *dna2*	0.2% (1/480)	0.007%±0.004%	1					
pol4 pif1-m2 *dna2*	0% (0/160)	0.014%±0.005%	N/A					
spt3 pif1-m2 *dna2*	6.3% (47/746)	0.117%±0.055%	47	55	7	3.6%	18.2%	78.2%
fob1 pif1-m2 *dna2*	6.5% (107/1646)	0.077%±0.040%	105	118	11	9.3%	16.1%	74.6%
pif1 pol32 dna2	5.5% (29/527)	0.075%±0.023%	29	33	4	9.1%	18.2%	72.7%
slx1 pif1-m2 *dna2*	8.6% (26/304)	0.069%±0.009%	5					
mus81 pif1-m2 *dna2*	3.6% (16/448)	0.081%±0.013%	10					
mus81 pif1-m2 *dna2*E675A	5.0% (8/160)	N/D	N/D					
sgs1	0% (0/240)	0.088%±0.012%	N/A					
exo1	0% (0/240)	0.104%±0.023%	N/A					
yen1^{ON}	0% (0/160)	0.044%±0.005%	N/A					
yen1^{ON} *dna2*	13.9% (30/216)	0.122%±0.043%	30	45	11	13.3%	6.7%	80.0%
yen1^{ON} *pif1*-m2 *dna2*	12.0% (24/200)	0.075%±0.015%	20					
pif1 dna2 + HU	23.0% (70/304)	0.057%±0.026%	N/D					

Table 1b

Genotype	Number of insertions at *URA3* among 5-FOA resistant colonies (%)	# of sequenced events	# of inserts	# of complex events	transposon insertions (%)	rDNA fragment insertions (%)	other nuclear genome insertions (%)
rad51	3.8% (5/133)	4	4	0	100%	0%	0%
rad51 pif1-m2	2.6% (4/154)	3	3	0	100%	0%	0%
rad51 pif1-m2 *dna2*	32.5% (96/295)	21	23	2	4.3%	26.1%	69.6%

Table 1c

Genotype	Number of insertions at *LYS2* (%)	NHEJ efficiency**	# of sequenced events	# of inserts	# of complex events	transposon insertions (%)	rDNA fragment insertions (%)	other nuclear genome insertions (%)
WT	0% (0/240)	0.828%±0.043%	N/A					
pif1-m2 *dna2*	2.9% (7/240)	0.288%±0.058%	7	8	1	37.5%	12.5%	50%

a, *MAT***a** locus. **b**, *URA3* locus. **c**, *LYS2* locus. *, Number of independent insertions and number of colonies tested for presence of insertion are shown; **, data are mean ± s.d.; $n \geq 3$ and represents the number of independent experiments; ***, the number is shown only for mutants for which at least 25 cases were sequenced.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Sequence analysis of DSB repair survivors in wild type and indicated mutants that do not carry a large insertion

Mutation	WT	*pif1*-m2 *dna2*	*rad27*
CGCAACA**(+CA)**GTA	13.3 (6/45)	26.6 (17/64)	17.1 (7/41)
CGC**(+A)**AACAGTA	4.4 (2/45)	1.6 (1/64)	4.9 (2/41)
CGCAACA**(+ACA)**GTA	2.2 (1/45)	0	2.4 (1/41)
CGCAA**(+AA)**CAGTA	0	3.1 (2/64)	7.3 (3/41)
CGCAAC**(+C)**AGTA	0	0	2.4 (1/41)
CGCA**(+C)**ACAGTA	0	0	2.4 (1/41)
CGCA**(-ACA)**GTA	44.4 (20/45)	26.6 (17/64)	17.1 (7/41)
CGCAA**(-CA)**GTA	13.3 (6/45)	14.1 (9/64)	22.0 (9/41)
CGC**(-A)**ACAGTA	6.7 (3/45)	10.9 (7/64)	19.5 (8/41)
CGCAACAG**(-T)**A	2.2 (1/45)	0	0
C**(-GCA)**ACAGTA	2.2 (1/45)	0	0
CGCAAC**(-AGT)**A	2.2 (1/45)	0	0
CGCAAC**(-A)**GTA	0	1.6 (1/64)	0
CGC**(-AA)**CAGTA	0	1.6 (1/64)	0
CGCAA**(-CAG)**TA	0	3.1 (2/64)	0
CGCAA**(-C)**AGTA	0	0	2.4 (1/41)
CGCAA**(-CAGT+A)**A	2.2 (1/45)	0	0
CGCAA**(-CAG+A)**TA	2.2 (1/45)	0	0
C**T**CAACAGTA	0	1.6 (1/64)	0
GGCAACAGTA	0	1.6 (1/64)	0
> 4 bp deletion	4.4 (2/45)	7.8 (5/64)	2.4 (1/41)

© 2018 Springer Nature Limited. All rights reserved.

TOOLBOX

GENE DESIGN GOES AUTOMATIC

Computer-aided design tools for genetic circuitry are starting to power synthetic biology.

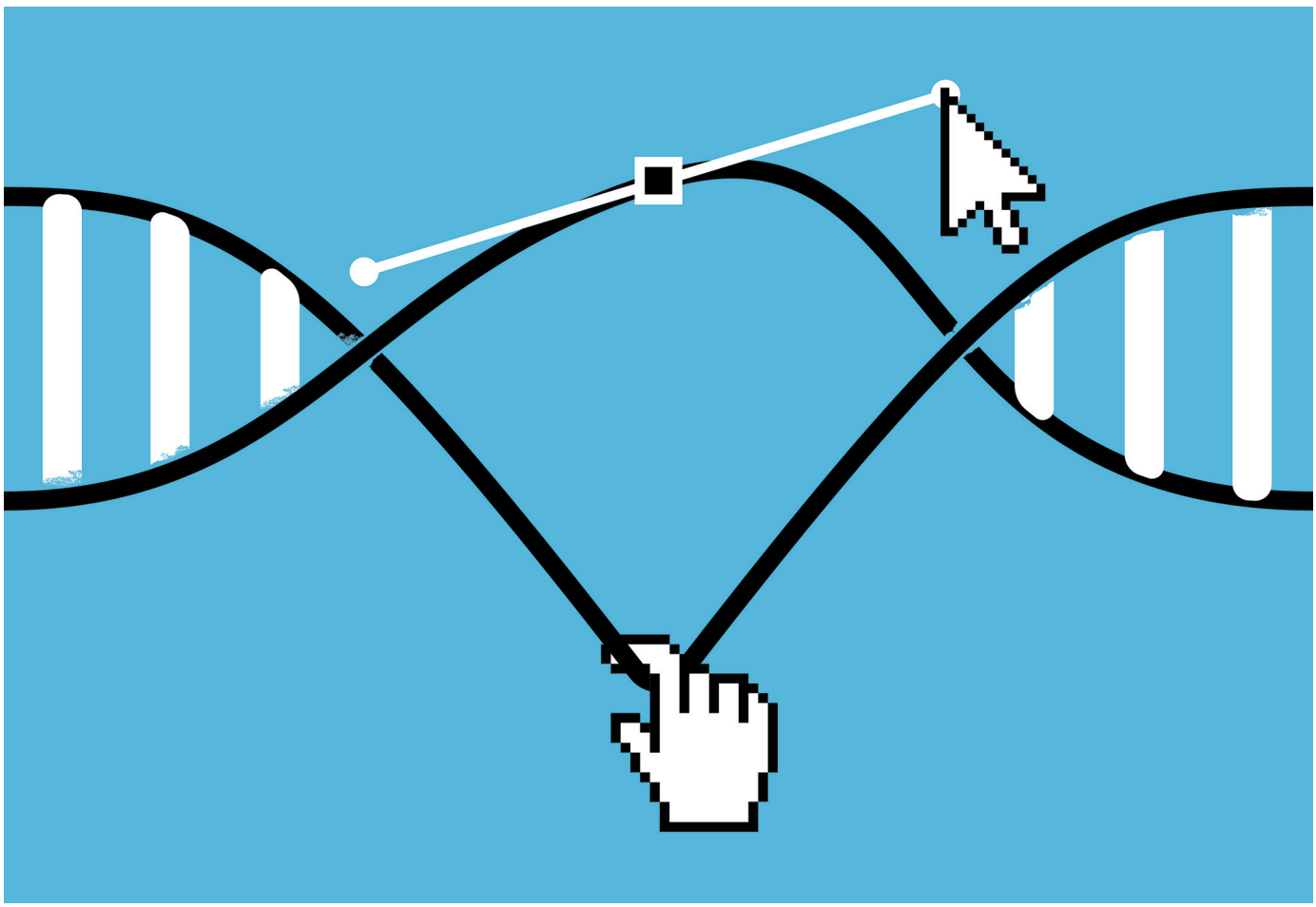


ILLUSTRATION BY THE PROJECT TWINS

BY ANNA NOWOGRODZKI

Your smartphone and laptop are made of electronic circuits. Genetic circuits, modelled on the electronic ones, are human-designed combinations of genetic components that interact to produce one or more proteins or RNA molecules, for example, in response to a given stimulus, such as a toxin. Under the right conditions, the circuit might be triggered to make "protein A, which then interacts with protein B to give outcome C", says David Riglar, a synthetic biologist at Harvard Medical School in Boston, Massachusetts. But until a decade or so ago, these two kinds of circuits were made in very different ways.

Electronics engineers design circuits using automated computer-aided design (CAD) tools. Genetic engineers, by contrast, have had to design biological circuits manually, and one at a time — a laborious, iterative and error-prone process. Computerized genetic-design tools are changing that. They automate the process by which researchers design complex genetic circuits that can program cells — especially bacteria and yeast — to carry out specific actions, such as activating a particular enzyme or churning out a certain protein. Synthetic biologists have used single-celled organisms in this way to produce drugs, biological sensors that include cells or antibodies, enzymes for use in industry, and more.

"Design tools for genetic circuits should greatly expand the accessibility of the kinds of genetic manipulations typically considered to be 'synthetic biology'," says Elizabeth Strychalski, a microbial engineer at the US National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland. Her group uses the genetic-design tools Cello and j5 to develop "living measurement systems", she says: cells that can act as sensors and ▶

© 2018 Springer Nature Limited. All rights reserved.

respond to their environment. "Those genetically engineered organisms then become tools in their own right, allowing anyone new ways to understand and control biology at the cellular scale."

According to Douglas Densmore, who heads the Cross-Disciplinary Integration of Design Automation Research lab at Boston University, such tools represent a fundamental shift in how genetic circuits are designed. Previously, he explains, genetic-circuit design was mostly a bespoke process. As a result, designs were difficult to share, improve and scale up. "It is not realistic to build an industry on artisanal approaches," he says. Although these "are great for early-stage research", they ultimately can't be transferred to a large scale. That's where automation comes in. "Automation will begin the process of getting designs out of notebooks and into software," he says.

A growing collection of circuit-design tools suggests that the field agrees. Yet software development for synthetic biology is in flux. One tool, Genetic Constructor, was abruptly terminated by its parent company, CAD software developer Autodesk Research in San Rafael, California, in August. But researchers working in single-celled organisms still have access to some open-source or freely available tools, including Cello, j5 and another called iBioSim. They can use these tools to weave circuits into whole genomes or to design thousands of variants to examine different combinations of genes, enzymes or protein domains.

"CAD tools are absolutely required for the design of biological systems," says Andrew Hessel, a genomic futurist and chief executive of Humane Genomic in San Francisco, California.

GENETIC CAD

Densmore, who developed Cello, has a background in electronic design automation — and it shows. Researchers can direct Cello to design a genetic circuit that meets certain specifications without having to tell the software anything about how to actually build it, just like with electronic-design tools. Users instruct the software — available both as source code and as a web application — using Verilog, the same computer language that electronics engineers use to describe their silicon circuits. "You specify the function you want, not the way it is created," Densmore explains. For instance, users could ask Cello to design a genetic circuit that produces a protein when it senses the presence of two particular antibodies. The software would then work out which components must be put together to make that happen, and output the nucleic-acid sequences required to physically build it. Cello also predicts how well its circuits are likely to perform.

Densmore designed Cello in collaboration with the lab of synthetic biologist Christopher Voigt at the Massachusetts Institute of Technology in Cambridge, for use in the bacterium *Escherichia coli*. Now, they are jointly expanding the tool to work in yeast, he says. Densmore and Voigt are using Cello to design circuits that produce a small signalling molecule in response to the presence of two other molecules, and are working on circuits with memory that function in different ways depending on the order in which they sense the targets, says Densmore.

Unlike Cello, other automated tools including iBioSim, j5 and GenoCAD do not spit out predictions for how well genetic circuits will perform or whether they're correct. And they all require the user to know and input information about how the circuit will be structured.

A GENETIC GRAMMAR

GenoCAD, which is commercial but has an open-source version, provides rules that define which functional parts of DNA sequences can go together, treating the sequences like programming code. "DNA sequences have the same linguistic complexity as programming languages — there are rules that people need to respect," explains Jean Peccoud, founder and chief executive of GenoFAB in San Francisco, which developed the software as the foundation for a broader set of genetic-design tools and services. "It's a grammar. All those rules are formal representations of biological knowledge." And from them, the software can translate a circuit design into the sequence for a physical piece of DNA, from which the circuit can be built. (Cello is built on a similar set of rules: a language called Eugene, which Densmore developed.)

"Genetically engineered organisms then become tools in their own right."

Created by the Joint BioEnergy Institute in Emeryville, California, and licensed exclusively to TeselaGen in San Francisco, j5 allows researchers to design genetic circuits by dragging and dropping genetic control elements onto a canvas. "You lay down a series of symbols that say, 'I want a promoter here, I want a ribosome binding site,'" says Densmore. Users can select multiple components that they might want to test in a particular location, for instance to work out which combination produces the most robust output. "Then you use rules to say, 'Don't put part A with part B, but part C has to be after part D', and then it enumerates all the different combinations," says Densmore. Researchers at non-profit universities and institutes can use the software through a free TeselaGen account; the firm also offers commercial accounts.

No special skills are required to use automated DNA-design tools, but because they do call for detailed specification of elements, familiarity with computer programming helps. "I don't think the learning curve is too steep right now, even in the more sophisticated tools," says Hessel. "None of these tools are so sophisticated that they couldn't be learned in a few days." The hard part, Hessel says, is building and testing the resulting circuits. Peccoud says he can teach even molecular biologists who have no computer-science background to use GenoCAD in just a few hours.

For those who need help, the greater synthetic-biology community is probably the best place to start. "Researchers in this field are generally accessible and helpful," says Drew Tack, a microbial engineer who works with Strychalski at NIST. "I would encourage someone just starting to reach out for advice and take full advantage of the considerable online resources" such as the GitHub code repository, he says.

One unfilled niche involves tools that are accessible for non-experts, but powerful and scalable enough to handle millions of base pairs of DNA and many designs. Before it shut down in August — along with Autodesk's entire life-sciences team — Genetic Constructor did just that. The closure "was an internal strategic decision", says Eli Groban, a computational biologist who led project management and strategy for the Autodesk life-sciences group When the company announced by e-mail that Genetic Constructor was ending, Groban says, it got replies from individual research groups asking them to keep it going. The tool's user interface was designed to be more accessible to the wider community of biologists than are tools aimed just at synthetic biologists, he says. "The gaps that Genetic Constructor wanted to fix still apply."

The use of genetic circuit-design tools is increasing among synthetic biologists, says Strychalski, albeit slowly. Groban says that the problem is one of economics. "In the academic community, there's this hesitation to pay for software. No one really does that cost–benefit analysis" that it might be cheaper to spend even tens of thousands of dollars on paid software than to get graduate students to spend significant amounts of time building their own version, or designing circuits in the old-fashioned way.

Right now, "most biologists don't work at scale", says Hessel, but that could be changing. He tells students that in their careers, they will operate on a vastly larger scale than their current lab work, managing liquid-handling robots and testing much bigger data sets of genetic variants. Automated genetic-design tools might well be required to make that happen. ■

Anna Nowogrodzki *is a science writer based in Boston, Massachusetts.*

CORRECTION

The Technology Feature 'The clinical code-breakers' (*Nature* **562**, 291–293; 2018) gave the wrong affiliation for Jenny Taylor. She is co-director of the translational genomic medicine programme at the Wellcome Trust Centre for Human Genetics.

© 2018 Springer Nature Limited. All rights reserved.

TRAINING Why undergrads in the lab benefit the whole team **go.nature.com/2pisr0d**

HIRING A postdoc disadvantage to finding non-academic jobs **go.nature.com/2syaacg**

SHARE YOUR STORY Contact us at **naturecareerseditor@nature.com**



RSSL

The sensory-evaluation team in action at UK testing lab Reading Scientific Services Ltd.

GASTRONOMY

Science you can taste

Food-industry scientists find diverse roles, from mediating public-health scares to perfecting meatless burgers.

BY DAVID PAYNE

In July 2016, a burger produced from fermented, genetically engineered yeast made its US restaurant debut. Its producers promised a taste, texture, nutritional value and sizzle similar to a meat patty.

Developed by Impossible Foods, a start-up founded in 2011 by biochemist and climate-change campaigner Patrick Brown, the burger's key ingredient is an iron-rich molecule called haem. Haem gives meat its flavour, but it can also be derived from plants: Impossible Foods found a way to extract it from soya-plant roots rather than ground beef.

The firm now has more than 300 employees and produces almost 500,000 kilograms of plant-based meat each month from its manufacturing site in Oakland, California.

The food industry is a complex and diverse sector, with companies ranging from start-ups to global multinationals. Its scientists include everyone from lab-based technicians who focus on foreign-body and microbiological-safety analysis to chemists, engineers, data analysts, social scientists and psychologists all working together on multidisciplinary research and development (R&D) teams.

In recent years, industry regulators have targeted certain strategies that some food companies use to market their products — advertising to children, and adding health claims and nutrition information on labels, for example — in response to public-health campaigns. As a result, some food scientists might face calls to explain the industry's role in tackling conditions such as obesity and type 2 diabetes.

Barbara Gallani, head of communication and engagement at the European Food Safety Authority in Parma, Italy, says that food science might not have the "glam factor" of, say, aviation or medical research, or the pace and pressures of academia, but "there's a lot of complex scientific work that goes on behind the scenes to produce everyday commodities".

In a 2016 TED talk, Brown, who left a "dream job" at Stanford University in California to start Impossible Foods, described his colleagues there as "brilliant, innovative, mission-driven scientists". Smita Shankar joined the company in 2013 as a protein chemist, and is now its director of research. Originally from India, Shankar completed a biochemistry PhD at Cornell University in Ithaca, New York, before moving to Hiten Madhani's yeast-genetics and molecular-biology lab at the University of California, San Francisco.

"At Cornell, I enjoyed teaching and doing good science in academia," she says. "But the San Francisco Bay Area exposed me to companies that were applying science to solve imminent problems, using microbial fermentation to make fuels, chemicals and other products."

Shankar leads a team of nine scientists and engineers who specialize in microbiology and industrial fermentation, collaborating closely with the group that extracts protein from the yeasts and adds them to the company's burger.

"Scaling up protein production from the lab to manufacturing is time consuming and expensive. Ultimately, the burger has to be delicious and affordable," she says.

A typical day for her team includes routine molecular biology and lab-based microbiology, alongside reading literature around yeast genetics. The researchers also hold a weekly group discussion to critically evaluate experimental design and discuss their data.

Other teams employ scientists who are trained in protein chemistry, polymers and textured materials, and flavour science. Their experiments aim to elucidate the properties of different animal-derived meats and determine how to recreate similar substances from plants.

Scientists at Impossible Foods are expected to have a PhD. "A PhD sets you up for working on hard problems. Also, you're not afraid of failing and trying new things," she says.

RISK COMMUNICATION

As with scientists in other sectors, food-industry researchers are often required to communicate complex science to the public, particularly in the event of a supply-chain ▶

© 2018 Springer Nature Limited. All rights reserved.

problem, food-disease outbreak or chemical or microbiological contamination.

Helen Munday, chief scientist at the Food and Drink Federation, a UK industry body, spent much of June 2018 explaining the role of carbon dioxide in food production to the media, after routine maintenance closed several European fertilizer plants and so triggered a shortage of the gas in the United Kingdom — right during the World Cup football tournament, prompting fears of a beer shortage.

Carbon dioxide is used to carbonate drinks and deliver beer from pressurized kegs to the glass, to stun animals before slaughter and as a preservative in packaged products.

"If you cannot stun pigs and chickens, you cannot slaughter them. They can't just be left on farms. There's a highly sensitive supply chain," says Munday, whose career includes seven years of managing pet-food innovation at a plant owned by global food giant Mars in Los Angeles, California.

Ian Noble's first overseas post as a UK food scientist was to the central highlands of Sri Lanka, where he helped to create and launch new tea-powder manufacturing processes. His work involved developing an understanding of polyphenols, a category of chemical that occurs naturally in plants, and in the case of tea, provides its distinctive flavour, and using that knowledge to concentrate the leaves into an affordable powder with a consistent colour and taste and a shelf life of up to six months.

Noble's PhD at the University of Reading, UK, was in biochemical engineering and colloid science. In food, whipped cream is an example of a colloid, containing one substance stabilized as a separate phase inside another, a gas inside a liquid. Milk is another: an emulsion of fat globules suspended in a liquid. Noble enjoys the challenge of combining physical chemistry and biochemical engineering, and "working to scale up production from millilitres in the lab to thousands of litres in manufacturing".

"How do you replace salt and still taste great? How do you replace sugar and keep bulk?"

He is now senior research, development and quality director at global snacks company Mondelēz International in Bournville, UK, the model village near Birmingham that was selected by the Cadbury family for its chocolate factory in 1879. Noble's fellow PhD colleagues include flavour chemists, toxicologists, microbiologists, computational modellers, mathematicians and theoretical physicists, who develop brands and products such as Oreos, Cadbury, Ritz and Toblerone.

Noble encourages food scientists to see themselves as commercial rather than academic researchers, whose understanding of science and technology as R&D professionals has earned them a place at the table where marketing and business decisions are made. "With food science, you are talking about fast-moving consumer goods and satisfying a marketplace demand for continuous new products," he says. "If I work on a product using a technological process that's already established in a factory, we can get it onto the shelf in a few months. Other sectors, including oil and gas technology, can take 25 years, so if you're working in those industries, you may have retired by the time it is introduced."

BRINGING HOME THE BREAD

Food-industry salaries vary widely across the world, with Germany topping the market.

Country	Median salary (US$)
Australia	91,603
Brazil	30,000
Canada	62,400
China	33,000
Germany	100,340
India	12,000
Mexico	31,000
New Zealand	82,800
United Kingdom	56,000

*Data included for countries with 10+ survey responses

A developed nation's food sector typically contributes around 5–10% of its overall economy, says Noble. In the United Kingdom, the agri-food sector was worth £113.2 billion (US$114.3 billion) in 2017. A 2017 report by the Committee for Economic Development of The Conference Board, a US public-policy nonprofit in Arlington, Virginia, says that the sector accounts for 5% of total US gross domestic product and 10% of total US employment (see go.nature.com/2khdbmi).

The Institute of Food Technologists in Chicago, Illinois, is an international membership organization representing food-industry scientists. Its 2017 pay survey found that the median annual salary of US food-industry scientists is US$92,000, and that female scientists in their 20s were paid similarly to their male colleagues, at $60,000 (see go.nature.com/2syyirl).

Flavourists came out on top, earning a median salary of $123,500. At the other end of the range were microbiologists, earning $65,000 on average (see 'Bringing home the bread'). Overall, 22% of survey respondents were educated to PhD level.

Jacinta George, managing director at Reading Scientific Services Ltd (RSSL), Mondelēz's UK testing lab, highlights vegan diets and protein-enriched snacks as two emerging industry trends driven by growing consumer demand.

As director of ingredient research between 2012 and 2015, she and her team focused on reducing sodium, sugar and saturated fat in its products. "There are big technology challenges," says George. "How do you replace salt and still taste great? How do you replace sugar and keep bulk, and have the right texture and microbial safety?"

Tim Lang, a social scientist who established City University London's Centre for Food Policy in 1994, says that the wholefood movement of the 1960s and 1970s paved the way for the current vegan trend. A long-time critic of many industry practices, Lang highlights other past trends, such as the rise of processed foods in Western diets, the market dominance of multinationals with huge advertising budgets and increased meat consumption globally. China, for example, where meat was once a rare luxury, now consumes 28% of the world's meat.

Lang says that scientists considering a food-industry career should think carefully about which branch of the sector to join. "There are wholefood firms, vegan firms, animal-welfare-obsessed firms, and there are firms that are the complete reverse of those," he says. Lang also recommends horticulture as a field in which scientists can develop biotechnology to support subsistence farmers.

For the past two years, Mondelēz has been developing a version of its Cadbury Dairy Milk chocolate bar that has 30% less sugar than the original. The product is due to launch in the United Kingdom and Ireland next year.

Some health campaigners urge the industry to go further in a bid to tackle the global obesity crisis. According to World Health Organization figures, more than 1.9 billion adults of 18 years and older were overweight as of 2016, and more than 650 million adults were obese. How do food-industry scientists address such criticism?

Noble highlights two documents: a foresight report by the UK Government Office for Science on the obesogenic environment, first published in 2007 (see go.nature.com/2zvxhnb), and the 2014 McKinsey Global Institute report, *How The World Could Better Fight Obesity* (see go.nature.com/2dsqdta).

Both reports, he argues, make clear that the challenges are multifaceted and that reformulation of food products — such as reducing fat, sugar and salt content — is only one of them. Lifestyle also plays a part.

"We do need to change the quality and quantity, and for many people, the availability, of food," Noble says. "But it's going to take a while to change our biology to reflect the dramatic changes in our modern lifestyles.

"Also, we may be looking at the first or second generation of people in developed countries who don't know how to cook. Food technology is not on the UK school curriculum. In the United Kingdom, I think, we've broken our relationship with food," Noble says. "You can either sit outside and point the finger, or you can choose to help fix it from the inside." ■

SOURCE: M. E. KUHN & M. MALOCHLEB *EARNINGS GROW, WAGE GAP SHRINKS* (INSTITUTE OF FOOD TECHNOLOGISTS, 2018).

David Payne *is chief Careers editor at* Nature.

CORRECTION

The Careers Feature 'Fathers in science' (*Nature* **563**, 725–727; 2018) gave the wrong name for Brian Cahill's wife. Her name is Lini, not Lina.

© 2018 Springer Nature Limited. All rights reserved.

COLD HEART

Communication is key.

BY VICTORIA DIXON



ILLUSTRATION BY JACEY

I taste ozoned rain, electrified metal and blood. It's cold. I'm close to hibernation. One tentacle throbs.

The spaceship lies in pieces and Imala is gone. I unbuckle my harness, yank exposed cords loose and fashion a sling, aching as I crawl from the wreckage.

Trees and grass burn despite the rain, and grey light reveals our wake of brutalized forest.

I cannot see my beloved, but send warmth to touch her.

There is nothing to touch.

Scorching despair ignites ship debris, and worse. Humans came to help. I try to quell my anguish and save them, but it's too late.

My rescuers ignite like walking saplings, shrieking. *What have I done?* They disintegrate. I can do nothing for our agony.

More of them hide until the fires die. Perhaps they hope I'll perish, but my people are born of heat. Cold is our anathema. I waver in and out of hibernation when they return in protective suits.

"Don't touch it!" a helmet-distorted voice says. They use machines to scoop up my body.

I do not blame them, but their distance is a cold anaesthesia. I wake in a laboratory.

"It's moving again," one man says. "Drop the temperature another degree." They heal my torn flesh while I sleep and mourn.

This is not kindness. It may be torture. I do not care. I have killed and deserve my fate.

They imprison me in a glass cell surrounded by tiers of oppressive black metal and unblinking lights. The door above is a series of metal panels and locks.

I wake from hibernation periods and one-sided studies with puncture wounds until at last, humans try to communicate. Our gestures are meaningless to one another. They bombard me with sounds that make my heart jump, but my people do not make sounds. We share warmth. Sometimes light.

However, fluorescent communications are involuntary, private conversations between family members. I cannot make myself change a single cell colour.

I try to share my warmth, but they must believe my grief was an attack. My temperature spikes and coolant flows into my cell. When I awake, those scientists are gone.

During my incarceration, 15 people have watched and studied me. Two left an impression. Of those, one remains. George, whose hair when he was younger appeared to combust into shocks of red and gold, never comes close. But long ago, he followed Halia like I used to drink Imala's warmth.

Halia — the only Terran who ever approached without fear.

Just once.

She stared at me, her head cocked. One lock of brown hair caressed her cheek. "It doesn't look healthy."

I had landed the year before. Their language was still strange. I will never speak as they do, but by then, I understood meanings and some nuances. I understood 'It'. Names encourage friendship and warmth; things I would never know again. So I thought.

"Do you need anything?" she asked.

Caring enough to ask was more sustenance than all their protein and vegetables.

"What is your name?"

She stepped close enough for me to tell her everything. I summoned warmth from my extremities.

George pulled her from the room, pointing out my elevated heart rate and temperature. White-faced caretakers sent me into hibernation for two days. When I awoke, George glared at me with red-rimmed eyes.

They fired her. This is one of their strange phrases. To me, this implies warmth and light, but to them? I hope it doesn't mean they killed her.

Now there's a bright yellow line on the floor where she turned back, and no one steps beyond it without placing me in hibernation.

I long for friendship. Warmth. Family. Years pass and George hates me more than ever. I understand this. I loathed myself after the crash. It was my idea to study humans — my fault Imala died. My fault Halia was fired.

"You're dead." George's face, once narrow, now overflows his white collar. His hair has thinned, but he remains the boy I remember. "They think you're dead. They won't rescue you."

I gather warmth. "*I know.*"

"No one else knows you exist." George steps on Halia's yellow line.

I form warmth into words and impress Halia's beauty and kindness on his thoughts. "*I loved her, too. I will never hurt you.*" Love is a warmth I know we can share. If he'll let me.

Water fills his eyes.

"*Did you hear me?*" He is so close. Like Halia in her extraordinary act of kindness. But as much as I love him, I know George is not kind. Not like her.

"You're dead and buried. No one's gonna find a body." He jabs the coolant button. I hibernate. When I wake, George is gone. I am not alone, but abandoned.

Days pass before they return. My heart rate and temperature ping warning sounds.

George *and* Halia. The years were kinder to her, but he still follows her with questions in his eyes.

I loved her too. I will never hurt you.

He'd heard me. Understood me in a mix of warmth and words that would have awed Imala.

The communications team enters. They connect new instruments to the temperature gauges. Joy and heat overflow. I fluoresce brilliant greens and vermilions. Blaze into colours invisible to the human eye.

My jailers protected themselves and I've never blamed them. Did they fear my anger, or that I might touch their cold hearts and reveal that I'm more than a thing?

My friends, my captors. They are all I have, but were never family.

Until today.

Halia takes George's hand and sobs. "He's beautiful." ■

Victoria Dixon *is obsessed with writing, culture, books, faith and family, although not in that order. She lives in Kansas, which is not monochrome, regardless of what fraudulent wizards might suggest.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

© 2018 Springer Nature Limited. All rights reserved.

nature OUTLOOK

GENE THERAPY

13 December 2018 / Vol 564 / Issue No 7735



Cover art: Sam Falconer

Editorial
Herb Brody, Richard Hodson, Elizabeth Batty, Lewis Packwood, Nick Haines

Art & Design
Mohamed Ashour, Wesley Fernandes, Kate Duncan

Production
Nick Bruni, Karl Smart, Ian Pope

Sponsorship
David Bagshaw, Jay Berfas, Anushree Roy

Marketing
Nicole Jackson

Project Manager
Rebecca Jones

Creative Director
Wojtek Urbanek

Publisher
Richard Hughes

Editorial Director
Stephen Pincock

Magazine Editor
Helen Pearson

Editor-in-Chief
Magdalena Skipper

Pharmaceuticals cannot always fix a malfunctioning human body. Sometimes the only way to treat what ails a person is to tinker with their genes: the blueprints for how biological systems are built and how they operate. Some researchers are using gene-editing techniques such as CRISPR to precisely alter DNA sequences. Others are genetically modifying immune cells to imbue them with the ability to fight cancer. And in the past couple of years, there has been a rapid acceleration in the development of a wide range of treatments in which disease-causing genes are replaced in their entirety.

This Outlook therefore focuses on the rich assortment of research in which new genes are introduced into a person, usually by means of a viral vector (see page S18). Successful animal experiments indicate that human genetic disorders could one day be repaired in the womb, so that a baby might enter the world disease-free (S6). And a number of health issues that have proved difficult or impossible to remedy — such as sickle-cell disease (S12), epilepsy (S10) and certain intractable skin conditions (S14) — might be excellent targets for gene therapy.

But gene therapy need not be limited to diseases that originate from genetic abnormalities. It might be possible to treat some viral infections with DNA, by using it to prompt the body into creating just the right monoclonal antibodies to ward off invading pathogens (S16).

Gene therapy remains an expensive medical path, however. Moving it out of the laboratory and into the clinic will require innovative pricing schemes (S23) and regulatory policies (S20). Along the way, clinicians, patients and policymakers will grapple with tricky ethical questions (S9).

We are pleased to acknowledge the financial support of Pfizer Inc. in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

Herb Brody
Chief supplements editor

CONTENTS

S6 NEONATOLOGY
The fix is *in utero*
Hereditary diseases healed prenatally

S9 PERSPECTIVE
A genetically augmented future
Ellen Wright Clayton on medical ethics

S10 NEUROLOGY
Repairs for a runaway brain
A way to suppress epileptic seizures

S12 BLOOD DISEASE
Medicine is in the blood
Fixing the gene in sickle-cell disease

S14 DERMATOLOGY
Under the skin
Epidermal cells extracted and repaired

S16 IMMUNOLOGY
A genetic shortcut
How to make an antibody factory

S18 THERAPEUTICS
Special delivery
Making viral vectors more efficient

S20 POLICY
Regulating a revolution
Health authorities tackle gene therapy

S23 PERSPECTIVE
Access and affordability for all
Michael Sherman on value-based deals

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, *Nature* **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2018).

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Gene therapy* supplement can be found at www.nature.com/collections/gene-therapy-outlook
It features all newly commissioned content as well as a selection of relevant previously published material that is made freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
Site licences (www.nature.com/libraries/site_licences): Americas, institutions@natureny.com; Asia-Pacific, http://nature.asia/jp-contact; Australia/New Zealand, nature@macmillan.com.au; Europe/ROW, institutions@nature.com; India, npgindia@nature.com. Personal subscriptions: UK/Europe/ROW, subscriptions@nature.com; USA/Canada/Latin America, subscriptions@us.nature.com; Japan, http://nature.asia/jp-contact; China, http://nature.asia/china-subscribe; Korea, www.natureasia.com/ko-kr/subscribe.

CUSTOMER SERVICES
Feedback@nature.com

Copyright © 2018 Springer Nature Ltd. All rights reserved.

© 2018 Springer Nature Limited. All rights reserved.

NEONATOLOGY

The fix is *in utero*

Some genetic diseases cause damage even before a child is born. The answer to these devastating conditions could lie in gene therapy delivered while the baby is still in the womb.



JOHN FEDELE/GETTY

Hereditary disorders that are discovered during prenatal scans could one day be cured before birth.

BY SARAH DEWEERDT

In July, an international team of researchers reported that they had used gene therapy to correct a fatal brain disorder in mice — before the mice were even born[1].

The mice had a defect in a gene known as *GBA*, which encodes an enzyme responsible for breaking down a fatty molecule called glucocerebroside. Without the enzyme, glucocerebroside builds up in the brain, causing irreversible brain damage. The mice typically die within about 14 days of birth.

The mice model a condition in humans called acute neuronopathic Gaucher's disease. Children born with this genetic mutation rarely live past the age of two.

In the study, researchers injected a virus bearing an intact copy of the *GBA* gene into the brains of fetal mice about mid-way through gestation. The treated mice were born normally, and lived for at least 18 weeks with little evidence of brain pathology. "You're talking about prolonging life significantly," says Jerry Chan, a fetal-medicine specialist at Duke-NUS Medical School in Singapore and an author of the study.

The researchers also administered the gene therapy to healthy macaque fetuses, and showed that it could transform brain tissue without serious side effects in an animal model that more closely approximates the body size and pregnancy physiology of humans.

"What we were trying to do is show the best possible experiments that would justify, if there ever was, a path to human clinical translation," says study leader Simon Waddington, a gene-therapy researcher at University College London.

Other researchers in the small field of prenatal gene therapy see the research as a leap forward, and say it provides the strongest evidence yet that the approach could be feasible in humans. "The combination of those two aspects of the study made it very, very exciting," says Bill Peranteau, a fetal surgeon at the Children's Hospital of Philadelphia in Pennsylvania.

The technical challenges, safety concerns and ethical issues of prenatal gene therapy are substantial. But this approach is more than just hotshot medicine. It could be the best way to treat a select group of devastating genetic diseases — and perhaps the only way to achieve a lasting cure.

EARLY ADVANTAGES

Acute neuronopathic Gaucher's disease is one of the best candidates for treatment with prenatal gene therapy. That's because the build-up of glucocerebroside begins in the fetus. In the absence of any intervention, irreversible brain damage can occur even before birth. "The

© 2018 Springer Nature Limited. All rights reserved.

main advantage is preventing the damage from occurring in the first place," Waddington says.

With other genetic diseases, the effects might not begin until sometime in infancy or early childhood. But even then, prenatal gene therapy might be more effective or efficient than waiting until after birth. "You are trying to take advantage of the normal developmental properties of the fetus to increase the efficiency and the likelihood of success of the treatment," says Peranteau, who is working on animal studies of prenatal gene therapy for metabolic diseases affecting the liver.

Before birth, the blood–brain barrier that prevents many molecules from crossing from the bloodstream into brain tissue is immature, a situation that eases delivery of genes to the central nervous system. In a 2011 paper[2], Waddington and his colleagues showed that a gene-therapy vector called AAV2/9 reaches nerve cells in the brain much more reliably in fetal mice than in those already born.

Another advantage of prenatal intervention is that the immune system is still immature. Therefore, the packaging used to deliver gene therapy — whether a virus or another type of vector — might be less likely to cause an adverse reaction. Also, the body develops immune tolerance to the vector, so if a gene therapy 'booster shot' needs to be administered later in life, it is more likely to succeed. The immune system will also accept the normal protein encoded by the gene therapy, rather than destroying it — as has sometimes been seen with postnatal gene therapy and protein-replacement therapies.

In addition, rapid fetal growth and development means more bang for the gene-therapy buck. At any given time, a large proportion of cells in the fetus is actively dividing. That yields a greater likelihood of the vector integrating into the genome. The population of corrected cells will continue to expand throughout gestation. Furthermore, to effect a cure, it is important to get replacement genes into stem cells or progenitor cells — and these long-lived cells are more abundant and more accessible before birth.

Finally, a 20-week fetus weighs roughly 300 grams, whereas a newborn tips the scales at around 3.5 kilograms. That small size translates directly into a higher therapeutic effect from a given dose of treatment. That's a big advantage because gene-therapy products can be expensive and laborious to produce.

A RISKY BUSINESS

But the fetal time period also poses unique challenges. Any prenatal intervention is complex because it affects two people — the mother and the fetus. "You've always got to take both into consideration, and you've also got to think about the future children of the mother herself," says Anna David, a fetal-medicine specialist and gene-therapy researcher at University College London.

Generally, the delivery of prenatal gene therapy is fairly straightforward. It involves injecting the treatment into an umbilical blood vessel, the amniotic fluid or occasionally directly into fetal tissue — often with the guidance of an ultrasound probe. The techniques are similar to well-established methods used in amniocentesis, chorionic-villus sampling or umbilical-vein blood transfusion.

"The procedures themselves are relatively safe," says David. Still, they do come with a small risk of infection, preterm labour and loss of the pregnancy. All in all, she says, "it's going to be a lot safer, probably, to treat it after the baby is born when you've got the baby and you're not risking the mother".

Then there are the usual risks involved in gene therapy, such as the potential for the vector to provoke an immune reaction in the patient, or incorporate into the genome in a location where it could trigger cancer. Some of these risks are magnified in the prenatal setting. For example, if the gene therapy gets into the mother's bloodstream, it could cause a dangerous immune reaction in her body or even be incorporated into her cells.

In the fetus, especially if given early in development, the gene therapy could alter germ cells that will eventually develop into eggs and sperm, causing changes that could be passed down to eventual offspring — a possibility that many scientists consider ethically problematic. The therapy might also disrupt normal body-system development by triggering the expression of genes in an inappropriate place or at an inappropriate time. That could potentially cause lasting effects, such as organ malformation.

Parents facing an *in utero* diagnosis of a serious genetic condition must often decide whether to raise a child with a lifelong disability or terminate the pregnancy. The appeal of prenatal gene therapy is that it offers a potential third path. But these treatments also raise the stakes: what if the gene therapy doesn't work, leaving parents with a seriously ill child they weren't prepared for and would not have chosen to raise? Similarly, a gene therapy that is only partially effective could turn a disease that previously would have been fatal in infancy into one that results in long-term disability — so it could actually increase suffering for the patient and family.

As a result of such concerns, researchers are cautious about the prospect of attempting prenatal gene therapy in humans. "If there is an adequate treatment for something after birth, that is the way to go," Peranteau says.

ORIGIN STORY

Even so, scientists have been thinking about prenatal gene therapy for nearly as long as they have been working on postnatal gene therapy. The first proof-of-concept studies[3] in animal models, showing that a gene could be introduced into an organism before birth, were published in 1995 — just a couple of years after the first human gene-therapy trial.

Often, scientists have looked to the prenatal window not just for the opportunity to treat diseases that begin before birth, but as a way around some of the difficulties of postnatal gene therapy. Charles Coutelle at Imperial College London, says that what prompted him to enter the field in the mid-1990s was, "to be quite frank, frustration with the efficiency of gene therapy at the time".

Coutelle had been involved in one of the first human trials of gene therapy for cystic fibrosis, a genetic disorder that affects the lung and other organ systems. It was difficult to deliver gene therapy to the lungs of people with cystic fibrosis because even in young children, the airways were full of viscous mucus and scar tissue; immune-system dysfunction also presented a hurdle. Coutelle thought it might be easier to correct cystic fibrosis *in utero*, when amniotic fluid moves freely in and out of the lungs.

"You are trying to take advantage of the normal developmental properties of the fetus."

Coutelle and his team spent several years perfecting fetal transfer techniques in mouse models, as well as working out which vectors would be best to use prenatally against cystic fibrosis or other serious diseases. The first big success — and an achievement that remains significant today — came in 2004. That year, a group including Coutelle and Waddington corrected the bleeding disorder haemophilia B in prenatal mice by injecting them with a virus bearing an intact copy of factor IX, a protein involved in blood clotting[4].

But the team soon had to switch gears. One vector used in the haemophilia work yielded only a temporary cure; another produced more lasting results but led to an increased risk of liver tumours. More importantly, the development of postnatal gene therapy for haemophilia had taken a sudden leap forward. "Once you have an established postnatal gene therapy there's no point in doing it prenatally. Or you have to have good reasons for doing it," Coutelle says.

A SURFEIT OF TARGETS

Waddington decided to look for a more challenging target disease that causes more severe effects earlier on, which led him to Gaucher's disease. But that is just one of a fairly broad array of metabolic disorders, including Tay–Sachs disease, Niemann–Pick disease and mucopolysaccharidosis, that cause *in utero* damage and could therefore be good targets for prenatal gene therapy.

Other researchers argue that haemophilia remains a good prenatal target. Researchers led by Graça Almeida-Porada and Christopher Porada at Wake Forest University in Winston-Salem, North Carolina, are working with a sheep model of haemophilia A. This form of haemophilia accounts for about 80% of haemophilia cases in humans, but has proven

© 2018 Springer Nature Limited. All rights reserved.

much more difficult to address with postnatal gene therapy than has haemophilia B.

One major issue is that the protein involved in haemophilia A — factor VIII — is highly immunogenic. Many people with a severe form of haemophilia A develop antibodies against factor VIII, which makes replacement therapy more costly and complicated, says Almeida-Porada. "The goal of going prior to birth is that you would induce tolerance to the protein — these patients would never develop an immune response," she explains. The team aims to cure haemophilia A in fetal sheep by collecting stem cells from the amniotic fluid, correcting the factor VIII gene and infusing the cells back into the fetus.

Studies of prenatal gene therapy in animal models are a dance between practicality and possibility. They depend on the availability of animal models for a given disease, and are shaped by the pace of advances in postnatal therapy or other experimental treatments, such as *in utero* stem-cell therapy or bone-marrow transplantation.

In June, researchers at Yale University in New Haven, Connecticut, reported that they had corrected the inherited blood disorder β-thalassaemia in fetal mice[5]. The disease is caused by mutations in the β-globin gene, which encodes a subunit of haemoglobin, the oxygen-carrying protein found in red blood cells. In β-thalassaemia, haemoglobin is less able to carry oxygen, leading to fatigue, growth stunting and damage to organs.

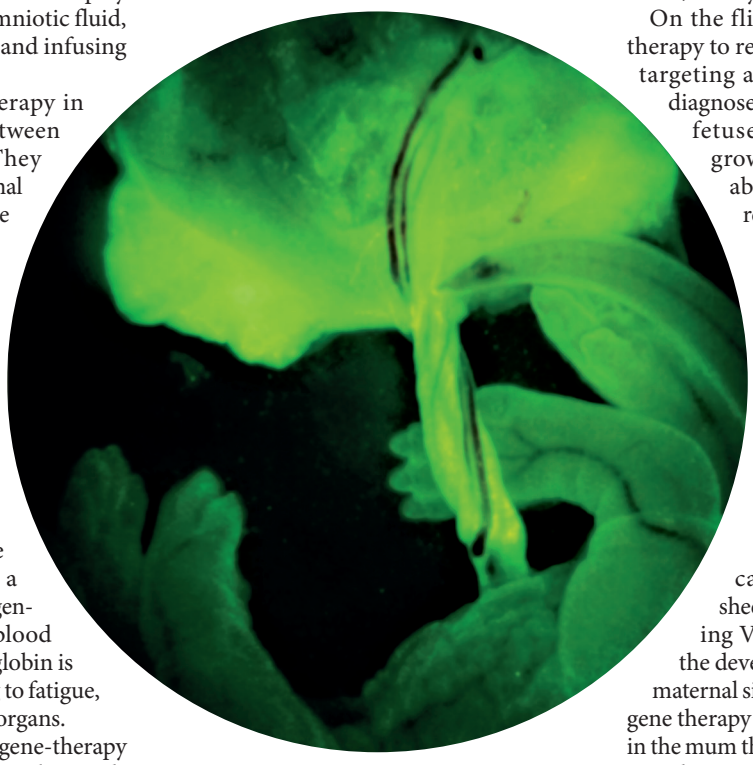
In the study, researchers used gene-therapy delivery vehicles called peptide nucleic acids (PNAs). PNAs are particles consisting of a biocompatible polymer surrounding an intact copy of the β-globin gene. "*In utero* injection of these molecules with a single injection was effective to achieve a phenotypic correction in the mice after birth," says study author Peter Glazer, a radiation oncologist and geneticist at Yale.

The PNAs make use of a cell's own DNA-repair mechanisms to incorporate the correct copy of the β-globin gene into the genome, potentially sidestepping some of the safety issues associated with gene-therapy delivery by viruses. And, crucially, the approach might be more effective prenatally than it is after birth. "In the developing fetus, the cells are more amenable to gene editing," Glazer says. "The DNA-repair capacity of the cells is revved up" because cells are dividing so rapidly, his team's data suggest.

Glazer envisions PNA-based gene therapy for thalassaemia or sickle-cell disease (another inherited blood disorder) being tried first in children, then infants and finally *in utero*. But how quickly this might happen is not clear. "For thalassaemia, a stem-cell approach is probably going to reach clinical practice much faster," says Chan. The safety of stem-cell or bone-marrow transplantation is better established than that of gene therapy, he says.

A BOON FOR RESEARCH

But even if prenatal gene therapy doesn't reach the clinic, it could still be useful as a research tool. That's already the case with cystic fibrosis, says Marianne Carlon, a gene-therapy researcher at the Catholic University of Leuven in Belgium.



ADELE RICCIARDI/DAVID STITTELMAN LAB/YALE

Fluorescent nanoparticles reveal a mouse fetus, umbilical cord and placenta.

Carlon and her colleagues have found that gene-therapy vectors can distribute more evenly through the lungs of fetal pigs than through the lungs of newborn pigs. The question is whether such even distribution is necessary or whether just reaching the large- and medium-sized airways is sufficient to prevent the lung damage in cystic fibrosis. *In utero* studies in animal models could also help to resolve questions about which cell types in the airways need to be targeted for gene therapy to be effective in cystic fibrosis.

"We would rather start in a neonatal setting" for attempting gene therapy on cystic fibrosis, Carlon says. Then, she adds, it would make sense to "move towards a fetal setting if you really see that you have difficulties targeting the right cell".

One reason that prenatal gene therapy for cystic fibrosis is not likely to be practical is that *in utero* screening for the disease is not widespread. As a result, the diagnosis is rarely made until after birth. "Without a prenatal diagnosis there is no prenatal gene therapy," Coutelle says.

Clinicians would need to be able not only to detect a disease before birth, but also to confidently predict that its severity would be sufficient to warrant gene therapy. These are complex questions that aren't fully resolved for all the prenatal target disorders. However, if there is no prenatal treatment for a disease, there might be little point in identifying it *in utero*.

Waddington's attitude is simply to bypass this catch-22 situation. "We'll develop the cures, and then that justifies doing the diagnoses," he says.

On the flip side, the first prenatal gene therapy to reach human trials might be one targeting a condition that is exclusively diagnosed *in utero* because it only affects fetuses before birth. Intrauterine growth restriction (IUGR) affects about 3% of all pregnancies and results in babies with dangerously low birth weight.

Unlike other prenatal gene therapy targets, IUGR is not a single-gene disorder. It occurs when, for unknown reasons, the normal remodelling of uterine arteries during pregnancy does not occur. That leaves the placenta and developing fetus starved of blood and nutrients.

David has shown that IUGR can be alleviated — at least in sheep — by delivering a gene encoding VEGF, a protein that stimulates the development of blood vessels, to the maternal side of the placenta[6]. "We're giving gene therapy to the mum, to treat a condition in the mum that causes a problem in the fetus," David says.

VEGF is expressed for only about a week, but that's long enough to trigger expansion of the placental vasculature. A similar approach has been used to stimulate the growth of blood vessels in the heart and neck, so the therapy, known as therapeutic angiogenesis, is well established postnatally. David has applied for regulatory and ethical approval to conduct a trial of the therapy in pregnant women.

"It's a major cause of cardiovascular disease and diabetes later in life," David says, referring to IUGR. "There's no treatment. And women want it, when you ask them. They're desperate to have a treatment." ■

Sarah DeWeerdt *is a science journalist in Seattle, Washington.*

1. Massaro, G. *et al. Nature Med.* **24**, 1317–1323 (2018).
2. Rahim, A. A. *et al. FASEB J.* **25**, 3505–3518 (2011).
3. Coutelle, C., Douar, A.-M., Colledge, W. H. & Froster, U. *Nature Med.* **1**, 864–866 (1995).
4. Waddington, S. N. *et al. Blood* **104**, 2714–2721 (2004).
5. Ricciardi, A. S. *et al. Nature Commun.* **9**, 2481 (2018).
6. Carr, D. J. *et al. Biol. Reprod.* **94**, 142 (2016).

© 2018 Springer Nature Limited. All rights reserved.

PERSPECTIVE



A genetically augmented future

Gene therapy could one day be used for bodily enhancement, creating an ethical minefield for physicians, says **Ellen Wright Clayton**.

The year is 2030. Gene therapy to insert the DNA sequence for dystrophin has been approved by regulators and is commonly used in children with Duchenne muscular dystrophy (DMD), a disorder linked to the X chromosome. Evidence shows that the intervention increases muscle mass in anyone who receives it. The treatment is widely available, but very expensive.

Alex, a slender adolescent, walks into a physician's office, accompanied by well-to-do parents. Alex does not have DMD, but wants to be stronger. Exercise is not providing enough benefits, and anabolic steroids have too many side effects. Alex is adamant about wanting dystrophin gene therapy and accurately cites its risks and benefits. Alex's parents are willing to pay for the treatment.

The cure for DMD described previously represents a cherished goal for gene therapy, and there is a lot of public support for fixing such heritable disorders in this way[1]. Yet Alex's request raises a host of questions.

We do not know why Alex wants to be stronger. Alex could have a milder form of muscular dystrophy or, if female, could be a carrier who experiences milder symptoms of DMD[2]. Alex might have some other cause of muscle weakness — or might want to be stronger for the sake of appearance, or to be more competitive in athletics. As is the case for many medical interventions, the potential uses of this therapy go beyond the specific disease for which it was developed. Possible applications range from treating milder disease to improving human characteristics — a continuum that could lack clear boundaries.

Let's assume that Alex does not have a diagnosed physical problem and wants the therapy simply to become stronger. The main debate about using medical interventions for such bodily enhancements focuses on the extent to which they give individuals an advantage over other people. A 2017 report by the US National Academies on gene editing in humans captures the debate well[1]. The authors summarize surveys that show that most people disapprove of using gene therapy to improve a person's physical and intellectual characteristics. The public tends to honour narratives of success based on personal diligence, or even accident of birth, over traits that can be purchased. This preference touches on a larger issue: the extent to which uses of gene therapy would exacerbate social inequality. If there is a widespread perception that this would be the result, society might try to limit its use to the few people who genuinely need it to treat their disease. Or there might be an effort to make such therapies available to all who want them.

Back to Alex in the world of 2030. Assuming that the US Food and Drug Administration's regulations are still the same, physicians would be free to use the approved DMD intervention for any purpose. After all, many medicines are legally prescribed for reasons that have nothing to do with their original indication. So what should happen? How hard should a physician try to understand the source of Alex's desire to be stronger?

Alex's wish might be a product of the social and cultural environment. The request might reflect issues with self-image. The desire to be stronger could reveal a psychological problem that needs to be resolved. Or a physician could conclude that Alex is suffering, thereby making the case for gene therapy more compelling. For example, medical and surgical interventions are sometimes prescribed to prevent or relieve psychological distress in children or young people who are abnormally short[3] or who have potentially stigmatizing physical features[4]. It is important to ensure that Alex understands and agrees to the therapy, but in the end, it can be hard to ascertain the source of a person's desire for a given treatment — especially if the person is an adolescent.

Are Alex's parents wrong to support their child's desire? Perhaps they are putting undue pressure on Alex. Perhaps they want to alleviate Alex's distress. Perhaps they are just indulgent. Society's default position is that parents should have the last say in such matters because they are assumed to care more for their children than does anyone else. Parents have a responsibility for shaping their children's future, creating opportunities and drilling into them all sorts of values. Parents are largely free to pursue their vision for their children's lives, unless those actions are illegal or constitute abuse or neglect.

CONCERNS ABOUT EQUITY SHOULD LEAD SOCIETY TO DEVELOP GUIDELINES FOR GENE THERAPY TO AVOID A NIGHTMARE FUTURE.

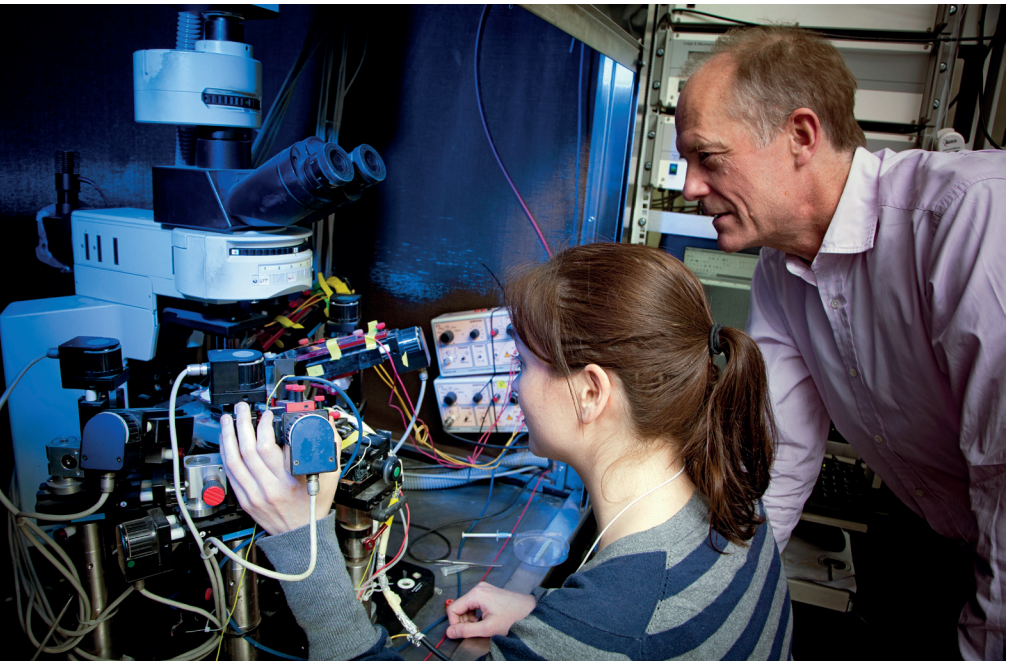
So what is the physician to do? Assuming that gene therapy for enhancement has not been outlawed, he or she can and should turn to medical ethics and the goals of medicine[5] for guidance. Respect for persons — a fundamental principle of medical ethics — would direct the physician to attempt to discover more about what is driving the patient and their parents' wishes, and to ensure that they understand what is at stake and that their expectations are realistic[6]. If the decision to proceed was made to relieve suffering, and with the adolescent's informed assent and the parents' permission, pursuing the goals of medicine would lead the physician to use the therapy to confer only traits within the normal range of human characteristics.

Ultimately, the ethics of enhancement are intertwined with views of fairness. Concerns about equity should lead society to develop guidelines for gene therapy to avoid a nightmare future in which a group of privileged people becomes stronger, smarter and more beautiful than the rest. But because drawing lines between treatment and enhancement is difficult, the more likely and more unsettling scenario is that physicians will be left to rely on their own ethical commitments to decide when to use gene therapy. ■

Ellen Wright Clayton *studies medical ethics and health policy at Vanderbilt University Medical Center in Nashville, Tennessee.*
e-mail: ellen.clayton@vanderbilt.edu

1. US National Academies of Sciences, Engineering and Medicine. *Human Genome Editing: Science, Ethics, and Governance* (National Academies, 2017).
2. Papa, R. *et al. Pediatr. Neurol.* **55**, 58–63 (2016).
3. Grimberg, A. *et al. Horm. Res. Paediatr.* **86**, 361–397 (2016).
4. Rohrich, R. J. & Cho, M.-J. *Plast. Reconstr. Surg.* **142**, 293e–302e (2018).
5. Allert, G. *et al. Hastings Cent. Rep.* **26**, S1–S27 (1996).
6. Grady, C. *N. Engl. J. Med.* **372**, 2172 (2015).

© 2018 Springer Nature Limited. All rights reserved.



DAVID BISHOP/UCL HEALTH CREATIVES

Elizabeth Nicholson and Dimitri Kullmann at University College London.

NEUROLOGY

Repairs for a runaway brain

Gene therapy could damp down epilepsy seizures in people for whom current drugs are ineffective.

BY LIAM DREW

The seizures of around one-third of people with epilepsy are resistant to available medicines — a statistic that haunts neurology. It has been this way for decades. The medicines have got better by becoming safer and causing fewer side effects. But still there are people for whom the drugs simply don't work — and for them, epilepsy can be ruinous.

"There's stigma; they can't drive; they have difficulty holding down jobs; they have difficulty maintaining relationships," says Dimitri Kullmann, a neurologist and neuroscientist at University College London (UCL).

Currently, the main hope for people with severe drug-resistant epilepsy is surgery. Someone whose seizures arise from a well-defined region of the brain might be offered an operation to remove that region. This is a drastic procedure, but not especially rare; it is carried out about 500 times every year in the United States.

Kullmann is hoping that gene therapy can make such surgery unnecessary. His group and others are investigating the potential benefits of introducing different genes into the brains of people with epilepsy, each one selected to quell the rampage of electrical activity that causes epileptic seizures. The most advanced projects are now being readied for clinical trials.

EXCITATION AND INHIBITION

Epilepsy comes in many forms. It is defined by the repeated occurrence of seizures — but these seizures can vary in their nature, intensity and frequency. And the disorder can arise from numerous causes, progress in different ways and affect distinct parts of the brain.

Crucially, epilepsy can either be focal, with seizures beginning in a specific brain region, or generalized, with seizures developing across wide spans of the brain. Focal epilepsy is more common, and it can be further subcategorized according to whether the seizures remain focal or spread to become generalized. There is also variation in the size of the seizure-generating focus and whether it is discrete, and therefore potentially removable, or enmeshed with vital brain tissue, and thus inoperable.

Brains essentially work by relaying electrical signals from neuron to neuron through the release of chemical neurotransmitters. Excitatory neurons release neurotransmitters that electrically stimulate neighbouring cells, whereas inhibitory neurons release neurotransmitters that suppress electrical activity. A seizure is a period of runaway electrical activity during which the normal balance between excitation and inhibition is lost. Current anti-seizure drugs either dampen excitatory mechanisms or boost inhibitory ones. But they do so indiscriminately, producing wide-ranging side effects by affecting neural circuits throughout the nervous system.

Current gene-therapy strategies, by contrast, use harmless viruses to introduce one or two therapeutic genes into the defined volume of tissue from which focal epilepsy emanates. "It is more personal, more targeted, and probably has fewer side effects because we treat the tissue that needs to be treated, instead of treating the whole body," says Merab Kokaia, a neuroscientist working on this approach at Lund University in Sweden. The strategies in development target focal epilepsy, but treating generalized epilepsy is a longer-term possibility.

RESTORING BALANCE

The brains of people with epilepsy contain increased amounts of neuropeptide Y (NPY), a chemical that certain neurons release when they are especially active. NPY acts on five receptors, Y1 to Y5, some of which are excitatory and some inhibitory. The levels of some of these receptors are also altered in epilepsy: notably, levels of Y2, which strongly inhibits neurotransmitter release, are higher. Overall, the accumulation of NPY and the altered levels of its receptors seem to represent an adaptive response — an intrinsic bid to hold back runaway brain activity.

In 2004, investigators used a viral vector to deliver the *NPY* gene into the brains of rats that had been manipulated to display a form of epileptic activity[1]. The resulting overexpression of NPY caused a reduction in seizure frequency. Other animal experiments also showed that overexpressing the neuropeptide galanin likewise suppressed seizures.

Kokaia, who was already working on NPY and epilepsy at the time, became interested in the therapeutic potential of this approach and started experimenting with introducing genes for neuropeptides, their receptors or both. He found that overexpressing NPY alone decreased seizure frequency, but simultaneously overexpressing it with the inhibitory Y2 receptor dramatically heightened the anti-seizure effect[2]. "What we are trying to do is boost the natural response of the brain by gene therapy," says Kokaia.

In 2015, Kokaia co-founded CombiGene in Lund, Sweden, to commercially develop this technique. In the past two years, CombiGene has confirmed the anti-seizure effects of the NPY–Y2 combination therapy, now called CG01, in further rodent models of epilepsy. And the company has successfully introduced the NPY and Y2 genes into brain tissue that was surgically removed from people with epilepsy.

Experiments using such tissue also ended

© 2018 Springer Nature Limited. All rights reserved.

ESTHER KROOK-MAGNUSON

interest in the second neuropeptide, galanin. Whereas NPY suppressed neurotransmission in human tissue, galanin did nothing — human neurons lack functional receptors for it.

Kullmann's move into the epilepsy field was serendipitous. His group was investigating the voltage-gated potassium channel Kv1.1 — a type of ion channel that electrically quiets neurons — as part of work on an entirely different neurological condition, episodic ataxia. The group made a virus that transferred the Kv1.1 gene into neurons. Because a neighbouring laboratory was routinely using rodent models of epilepsy, Kullmann and colleagues thought it might be worth testing Kv1.1 in these animals. The effect, published in 2012, was a dramatic reduction in seizure frequency[3]. After seeing this effect in three separate animal models, Kullmann and UCL colleague Stephanie Schorge developed a viral vector that introduces a modified Kv1.1 gene specifically into excitatory neurons, and does not integrate the gene into the cell's genome.

In principle, CG01 or Kv1.1 could provide long-term suppression of epileptic seizures following a single injection, with the genes continually generating products that calm the neurons in which they are expressed.

TRIGGERED ACTIVATION

Several alternative approaches are mainly based on converting widely used basic-research technologies into clinical tools. These approaches are more complicated, but hold potential advantages over CG01 or Kv1.1.

Opsins, for example, are membrane proteins that are activated by light, and the genes encoding them have been isolated from microorganisms. When illuminated, some types excite neurons, whereas others inhibit them. The big appeal of opsins is that they could remain inert in neurons when brain function is normal and only be called into action when needed.

Esther Krook-Magnuson, a neuroscientist at the University of Minnesota in Minneapolis, has shown that opsins can control seizures in rats[4]. Her team introduced inhibitory opsins into the rats' epileptic foci, then implanted seizure-detecting electrodes into their brains, along with fibre optics that light up to activate the opsins. An algorithm switched on the light when it detected the first signs of epileptic activity, quashing seizures early. Krook-Magnuson notes that implanting electrodes and light sources into humans would be less invasive than the current option of removing an area of brain.

However, this system requires a reliable seizure-detection method, an effective light-delivery technique and a way to get the right amount of virus into the right neurons. All three components will have to be optimized before the system has a chance of reaching the clinic.

The need to develop more than one technology can put off potential investors, says Kullmann. He has first-hand experience of this from trying to transform another research tool — DREADDs (designer receptors exclusively activated by designer drugs) — into a therapy. DREADDs are genetically engineered receptors that, like opsins, sit silently in neurons unless they are activated by a stimulus, but in this case, the stimulus is a drug rather than light.

Both Kullmann and Kokaia have found that inhibitory DREADDs can suppress seizures when the genes encoding them are inserted into the seizure foci of epileptic animals using viral vectors. If the therapy were translated to humans, people might take the activating drug regularly in a similar way to current epilepsy medicines — but with the advantage that the DREADDs would not inhibit brain tissue outside the region where the DREADD is situated. Alternatively, people might receive the drug automatically through an implanted, seizure-activated drug-delivery system, or simply take the drug when they feel the first indications of a seizure.

Kullmann is also exploring an ion channel that was originally identified in nematode worms. In nematodes, the glutamate-gated chloride (GluCl) channel is inhibitory and is activated by the neurotransmitter glutamate. But in mammals, glutamate is the main excitatory neurotransmitter that is responsible for driving excess activity during seizures, and none of its receptors is inhibitory.

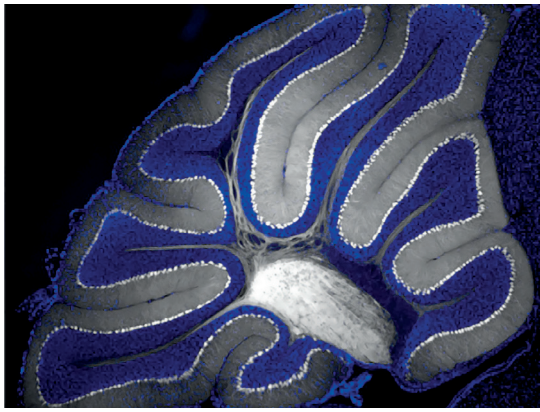
Kullmann and his colleague Andreas Lieb were interested in using an engineered version of the GluCl channel that is activated by a drug, but then they learnt that mutations in GluCl can change its glutamate sensitivity. If they picked a mutated channel that was insensitive to normal levels of glutamate, but activated by the high levels of glutamate that occur during seizures, they might have an appealing gene-therapy agent: an inhibitory ion channel that is ordinarily inactive but called into action during seizures. Early findings are encouraging: in two rat models, GluCl decreases seizure frequency[5].

PRIMED FOR CLINICAL TRIALS

In January, CombiGene partnered with the London-based incubator Cell and Gene Therapy Catapult to develop manufacturing processes for CG01 in preparation for clinical trials. And in April, Kullmann and Schorge received nearly £2 million (US$2.5 million) from the UK Medical Research Council to move the modified Kv1.1 virus towards the clinic.

Several technical hurdles remain, including scaling up the drug-delivery system: a human brain is around 700 times larger than the rat brains in which the viral vectors have been tested. But a major advantage of using NPY, Y2 and Kv1.1 is that they are derived from human genes — and therefore unlikely to evoke an immune response. By contrast, microbial opsins and GluCl from nematodes carry the risk of rejection by the immune system.

The hope is that gene-therapy treatments will be applicable to all drug-resistant focal epilepsies, including in people whose larger or awkwardly located foci make them ineligible for surgery, says CombiGene chief executive Jan Nilsson. And, more speculatively, if it is successful, gene therapy could potentially be adopted by some people instead of conventional drugs.



Brain cells could be manipulated using light.

But for the time being, CombiGene and Kullmann's team are planning safety and tolerability trials that will involve only people with drug-resistant epilepsy who are awaiting surgery. This is not because people in this group are the sole intended recipients of gene therapy — rather, they present a unique opportunity.

"What we are trying to do is boost the natural response of the brain by gene therapy"

The virus is likely to be given during presurgical investigations of the seizure locus, then allowed to enter neurons and deposit its genetic cargo while the patient spends weeks to months awaiting surgery. In phase I trials, surgeons will then almost certainly remove the focus. This procedure will allow researchers to carefully examine whether the gene delivery worked, and will also provide a fail-safe mechanism for excising genetically modified tissue should any safety issues arise.

The alternative is that people could opt out of surgery. If gene therapy is to be approved for epilepsy, numerous larger, more stringently controlled trials specifically designed to look at anti-seizure effects will be needed. But Kullmann allows himself to imagine a best-case scenario with the first exploratory trial. Someone who has stopped having seizures after the gene transfer, he says, might simply elect not to have surgery — entering a realm where their seizures are quelled not by conventional medication, but by DNA. ■

Liam Drew *is a freelance science writer in London.*

1. Richichi, C. *et al. J. Neurosci.* **24**, 3051–3059 (2004).
2. Ledri, L. N. *et al. Neurobiol. Dis.* **86**, 52–61 (2016).
3. Wykes, R. C. *et al. Sci. Transl. Med.* **4**, 161ra152 (2012).
4. Krook-Magnuson, E., Armstrong, C., Oijala, M. & Soltesz, I. *Nature Commun.* **4**, 1376 (2013).
5. Lieb, A. *et al. Nature Med.* **24**, 1324–1329 (2018).

© 2018 Springer Nature Limited. All rights reserved.



STEVE BABULJAK/UCSF

Six-year-old twins Tylee and Taleeke both have sickle-cell disease.

BLOOD DISEASE

Medicine is in the blood

Sickle-cell disease is an ideal target for gene therapy, but economic and social barriers to treatment are rife.

BY ANNA NOWOGRODZKI

Elliott Vichinsky estimates that at least 30% of his adult patients with sickle-cell disease die from preventable causes. Red blood cells are supposed to be shaped like concave discs, but in people with sickle-cell disease, a mutation in a single gene collapses them into a crescent shape. The pointy sickles catch on each other and clog blood vessels. They cut off oxygen to limbs. They cause kidney failure, hypertension, lung problems and strokes — along with bouts of excruciating pain.

These are common and treatable complications, so why the high death rate? Vichinsky attributes it to a lack of infrastructure, such as care centres, to properly monitor adults with sickle-cell disease. This is partly because the disease mainly affects low-income minorities and people in developing countries. "If they were tracked before," says Vichinsky, "they would not be dead."

Gene therapy might offer a cure for sickle-cell disease, and clinical trials are already under way. "In the long run I think it will be able to cure the disease," says Vichinsky, a haematologist and oncologist at the University of California, San Francisco (UCSF) Benioff Children's Hospital in Oakland. The approach is promising because just a single gene needs correcting: the one for the β-globin subunit of haemoglobin, the body's oxygen ferry. But Vichinsky is concerned that the same problems that make current care ineffective will also plague this gene-therapy treatment. As his patients attest, sickle-cell care is often inadequate for reasons that have little to do with scientific advancement and lots to do with economics and racism.

For people with sickle-cell disease in the United States, paying for the treatment could be a challenge: it involves such hefty upfront costs that insurers might not be able to cover the treatment, even if it saves them money in the long term.

The only current cure for sickle-cell disease is a bone-marrow transplant from a matched healthy donor. The stem cells that serve as blood-cell factories — haematopoetic stem cells — are removed from the donor's bone marrow or blood, then infused into the recipient. If the transplant works, the donor's stem cells churn out non-sickle-shaped red blood cells, curing the disease. Donors can be a sibling or someone unrelated with the same bone-marrow type, but less than one-third of people with sickle-cell disease can find a matched donor.

Gene therapy could provide a cure for many more people because it doesn't rely on a donor: instead, stem cells are harvested from the patient's own bone marrow. As a further benefit, gene therapy avoids conflict between the donor's and recipient's cells. After a bone-marrow transplant, doctors have to suppress the recipient's immune system to prevent it from attacking the transplant, which leaves the patient vulnerable to infection. Even then, the donor cells might attack the recipient's cells, resulting in graft-versus-host disease — the leading cause of death after a bone-marrow transplant. Gene therapy eliminates this concern.

GENE THERAPY ON TRIAL

Mark Walters, a paediatrician at UCSF Benioff, is working on two gene-therapy clinical trials. One by Bluebird Bio in Cambridge, Massachusetts, is in phase I/II, and one by Bioverativ in Waltham, Massachusetts, will start soon.

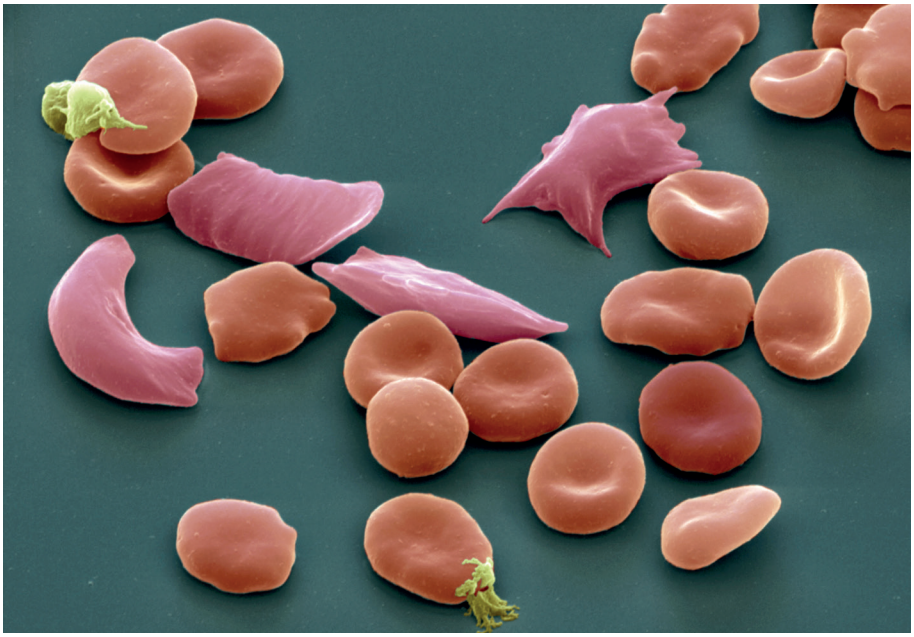
For the Bluebird Bio trial, Walters has enrolled two people so far, and plans to enrol four or five in all at his institution — a total of 50 people will be recruited across the United States. The trial is using the gene-therapy drug LentiGlobin BB305 to insert a healthy version of the β-globin gene into people's blood stem cells. With the gene, the stem cells will make normal red blood cells instead of sickle-shaped ones.

Stem cells are harvested from each person in the trial, and they receive blood transfusions every 3–4 weeks to reduce the percentage of sickle cells in their blood, says Walters. "We don't want patients having complications in the middle of the trial or leading up to it."

It takes about a month for the new gene to be inserted into the patients' stem cells. After being collected up, the cells are shipped overnight by plane to a central manufacturing location, where they spend several days just multiplying. Then scientists put the β-globin gene into the stem cells using LentiGlobin BB305, a vector made from a virus. After quality-control testing, the improved stem cells are frozen and shipped back to UCSF Benioff.

In the meantime, the patients receive four days of intensive chemotherapy to wipe out any remaining stem cells with the old, problematic version of the gene. The improved stem cells

© 2018 Springer Nature Limited. All rights reserved.



EYE OF SCIENCE/SPL

Normal red blood cells (red) compared with the elongated blood cells in sickle-cell disease (pink).

are then reinfused into the person around a day later, and their immune system regains its strength slowly. "It takes about three months to completely recover," says Walters.

A COSTLY ENDEAVOUR

The clinical trials will demonstrate whether gene therapy is effective at curing sickle-cell disease. But even if it is, the cost of treatment is likely to be very high. For example, voretigene neparvovec (Luxturna), a gene therapy for degenerative blindness, costs US$425,000 per eye. "We're looking upwards of $500,000 to $700,000" for sickle-cell gene therapy, spread over multiple years, says Stephanie Farnia, director of health policy and strategic relations at the American Society for Blood and Marrow Transplantation in Chicago, Illinois. And this is a disease for which more than 50% of patients in the United States rely on government health insurance such as Medicare and Medicaid.

In the long term, an expensive cure for sickle-cell disease would probably be cheaper than — and much more preferable to — dealing with 30–40 years of the disease's chronic, long-term effects. But even if the pharmaceutical company spreads the cost to insurers over 5–7 years, Farnia says, insurers, particularly government-funded ones, will probably not have sufficient capital to pay for everyone who wants the treatment. "The really tough part is these budgets do not have a lot of room in them for additional costs," Farnia says. It's like trying to pay for an entire 30-year mortgage in just five years, she says. "You're going to save a lot more money down the road, but can you come up with the money to do that?"

For a possible preview, Farnia suggests looking to chimeric antigen receptor T-cell (CAR-T) therapy — a type of immunotherapy that has shown promising results in treating certain types of cancer. US medical centres and hospitals are paying for CAR-T therapy up front to treat their patients, before knowing whether insurers will reimburse them for it. "And they have to hope they can figure out with payers that they get reimbursed for enough of that," Farnia says.

CHALLENGES AHEAD

There are other concerns with gene therapy as well. For one, more long-term monitoring is needed. The added gene slips in at random places in each stem cell's genome, so it has thousands of opportunities to land in the middle of another important gene. It could theoretically wind up in a gene that suppresses cancer. No one has yet observed a leukaemia caused by delivering treatments with the family of viral vectors that LentiGlobin BB305 belongs to, Walters says, but a stem cell is long-lived. "If you treat a child, it's going to be a source of blood for the next 50–60 years." No patients have been monitored for anywhere near that long after gene therapy.

Although gene therapy opens up bone-marrow transplants to more people than the one-third who have a suitable bone-marrow donor, it doesn't open it up to everyone. "It's still an intensive procedure," says Walters, particularly the high dose of chemotherapy that people receive before the stem cells are returned to their bodies. "Not everybody is well enough to go through it."

Recruiting for clinical trials might also be a problem. Current trials involve small numbers of people with sickle-cell disease, but if the treatments work, future trials will require many more participants. In the United States, sickle-cell disease is more common among black and Hispanic populations, and there is an ugly history of non-consensual medical research on black people, causing some to be wary of participating in clinical trials. And racial bias also gets in the way of treating the disease. "The hallmark of sickle-cell disease is pain, and it's excruciating pain. It's like putting a tourniquet on and depriving a limb of oxygen," says Walters. And unfortunately, doctors have been shown by multiple studies to be less likely to believe black people's claims to be in pain than white people's (see, for example, K. M. Hoffman *et al. Proc. Natl Acad. Sci. USA* **113**, 4296–4301; 2016).

> ***"Sickle-cell disease represents the best and worst of health care in the United States."***

Sickle-cell disease is a chronic condition. Management of chronic diseases isn't typically groundbreaking, and even among chronic diseases, sickle cell is typically neglected. "It's not received the attention or the national funding that it maybe should have received, because it's not as politically connected," says Walters.

Vichinsky argues that gene therapy should be part of a multidisciplinary programme that includes basic care, not a substitute for basic care. "We shouldn't push them into gene therapy just because there's no basic care available," he says. The US Centers for Disease Control and Prevention list 175 providers of paediatric care for sickle-cell disease in the United States, but only 44 providers of adult care. Vichinsky started his own adult programme because he had nowhere else to transfer his young patients when they became adults. "It has to do I think with money and ethnicity," he says.

Basic care for sickle-cell disease should be modelled on current programmes for cystic fibrosis or childhood cancer, says Vichinsky. He advocates that sickle-cell-disease medical centres should include multidisciplinary teams to monitor people for the degenerative effects of sickle cells across many different organ systems, such as the lungs, heart, kidneys, spleen and brain. That way, doctors could detect early warning signs of problems such as renal failure and hypertension.

He is optimistic, however, that sickle-cell gene therapy might act "as a kind of door opener to the field of gene therapy". There are a handful of gene-therapy drugs on the market, but sickle-cell disease's role as an early gene-therapy target, and the promise of that therapy, might attract interest in how best to care for people with this disease, and propel standards of care forward.

"Sickle-cell disease represents the best and worst of health care in the United States," Vichinsky says. Technologically advanced gene therapy is a hot research area, but not yet proven to work. Mundane chronic illness care is neglected, but it would save lives. "Most adults don't have access to multidisciplinary services," says Vichinsky. "I believe to some extent that gene therapy will actually stimulate the medical and scientific community to bring that to sickle cell." ■

Anna Nowogrodzki *is a science writer based in Boston, Massachusetts.*

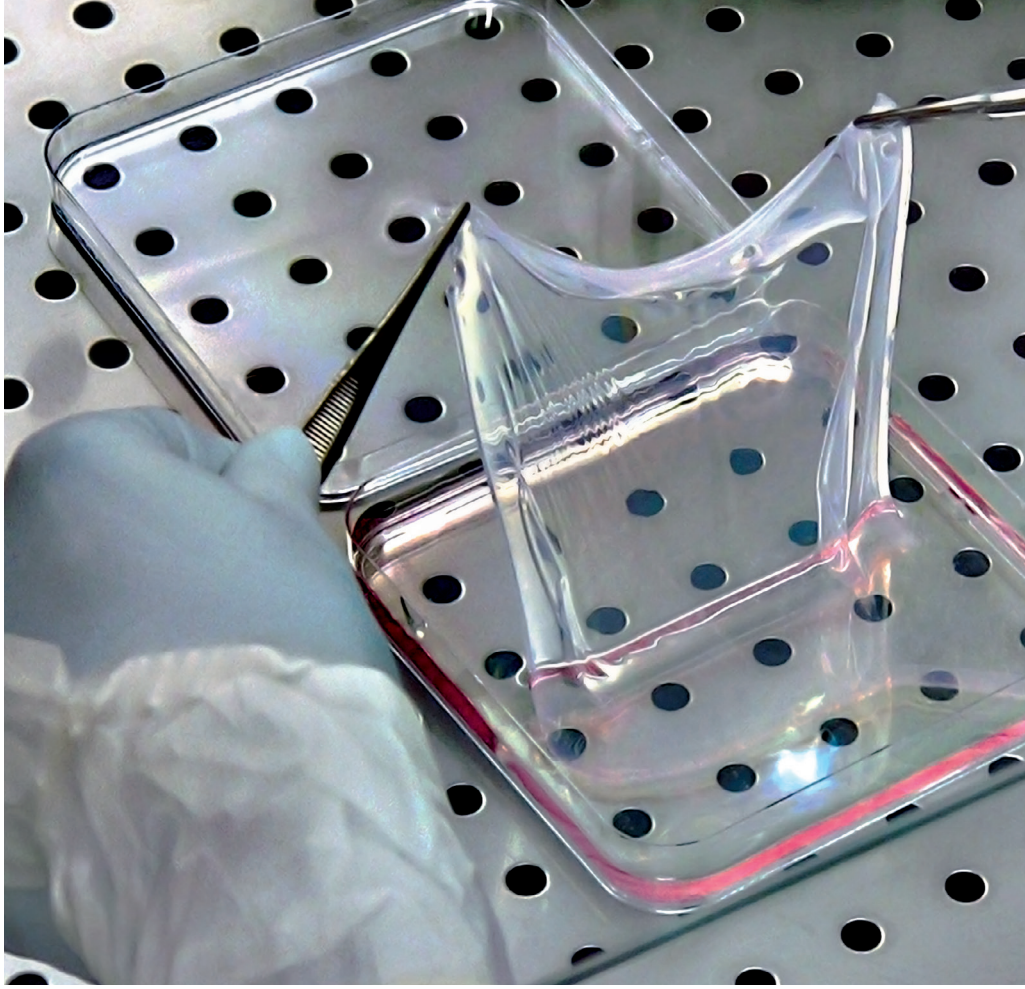
© 2018 Springer Nature Limited. All rights reserved.

DERMATOLOGY

Under the skin

The largest organ in the body is a prime target for gene therapy.

BY KAT ARNEY



CMR UNIMORE

It's not often that a figure in a scientific paper can make you wince with pain. But it's impossible to look at figure 1a in Michele De Luca's 2017 *Nature* paper and not feel a sympathetic twinge at the sight of a young boy, Hassan, covered from head to toe with red-raw wounds[1].

The son of Syrian refugees who fled to Germany, Hassan was born with junctional epidermolysis bullosa (JEB) — a condition caused by a genetic fault in one of three genes (*LAMA3*, *LAMB3* and *LAMC2*) encoding subunits of the laminin-332 protein, which binds the surface of the skin to the underlying layers. Affected children rapidly develop large, painful blisters over their skin and internal mucous membranes, which can easily become infected.

By 2015, when Hassan was seven, his skin was almost entirely destroyed and he was suffering from severe bacterial infections. Doctors at Ruhr University in Bochum, Germany, could offer only palliative care to relieve his suffering. But Hassan's father enquired about experimental treatments, and the doctors got in touch with De Luca at the University of Modena and Reggio Emilia, Italy, who was working on a radical skin therapy.

De Luca's research builds on the life-saving work of cell biologist Howard Green at the Massachusetts Institute of Technology in Cambridge. Green was the first to discover that sheets of skin cells could be grown in the laboratory, creating personalized skin grafts that avoid the problems of immune rejection. De Luca worked with Green at Harvard Medical School in Boston, Massachusetts, in the 1980s, and he later decided to develop Green's approach for treating genetic skin conditions by genetically modifying the skin cells to fix the disease-causing mutation.

"We've been using epidermal skin-cell cultures for many years to treat hundreds of patients, carrying out a lot of work on basic stem-cell biology as well as gaining clinical experience, so it was obvious to try and genetically modify these cells for treating rare skin diseases like JEB," De Luca says.

The idea of growing genetically modified skin for therapeutic use was first proposed in 1994 by dermatologist Gerald Krueger at the University of Utah in Salt Lake City[2], and De Luca and his team reported the results[3] from an initial small clinical trial of genetically modified skin grafting back in 2006.

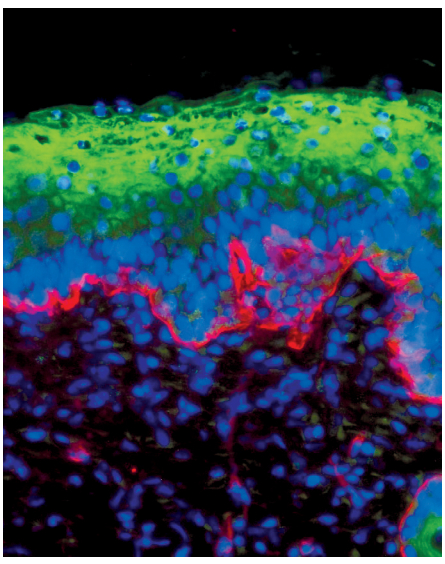
The recipient was a 36-year-old man with JEB caused by a *LAMB3* mutation. He was treated with nine small patches of skin that were grown from his own epidermal cells and modified with a viral vector expressing the missing gene. The grafts remained stable and healthy for more than a year, proving that the technique had the potential to provide long-term correction of the condition.

UNSCHEDULED INTERRUPTION

Despite this early success, De Luca's clinical work ground to a halt for nearly a decade owing to European Union legislation governing cell and gene therapies. "The regulations regarded our grafts as medical products, so they had to go through the same regulatory process," he sighs. "We had to stop all our activities, build up a compliant manufacturing facility and register the therapy — it was only in 2015 that we were finally able to start our trials again."

Luckily, this was just in time for Hassan. De Luca's team took a tiny unblistered skin sample from the child's groin, then carefully cultured the epidermal stem cells and modified them with a viral vector carrying a functional version of *LAMB3*. The next challenge was growing enough 12-centimetre-square sheets of modified cells for Ruhr University plastic surgeon Tobias Hirsch to wrap around the child's fragile body.

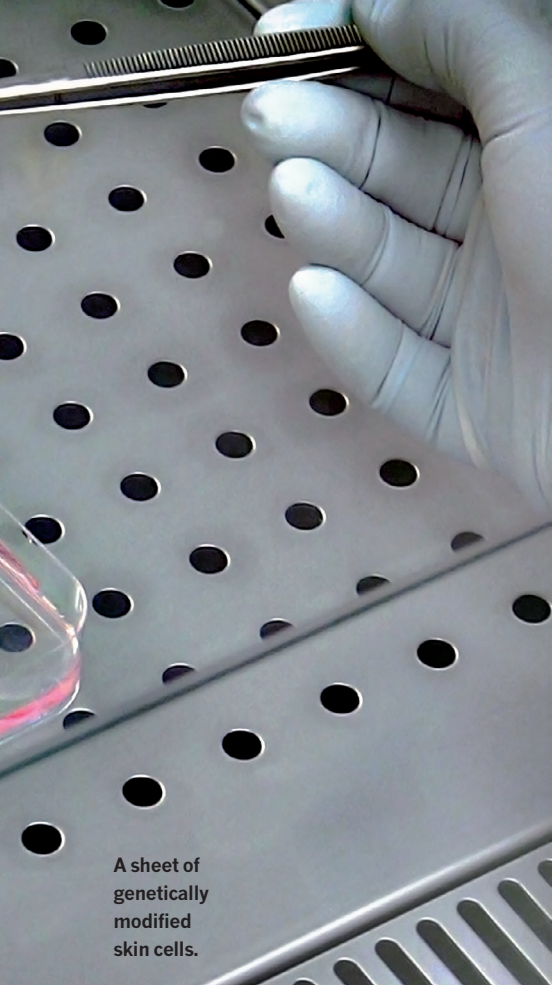
After two major operations to replace the skin on Hassan's limbs and torso, followed by some smaller procedures, around 80% of the entire epidermis had been replaced, making it the largest genetically modified graft performed to date. By the time the results were published in 2017, Hassan was like a different child, his raw blisters replaced with smooth, perfectly functional skin.



XIAOYANG WU/CYNTHIA LI

A skin graft with fluorescent staining.

© 2018 Springer Nature Limited. All rights reserved.



A sheet of genetically modified skin cells.

RUHR-UNIVERSITY BOCHUM

As well as detailing Hassan's progress, the paper[1] reveals why the treatment was a success. The skin is made up of many different types of cells, some that are short-lived and others that are much more persistent. The researchers showed that long-term grafting was only possible if the genetically modified cells were holoclones — a relatively rare type of immortal cell that can self-renew indefinitely. By adjusting the culture conditions, De Luca and his team were able to encourage the growth of holoclones, greatly increasing the chance that the resulting grafts would work.

"After three years, his skin is stable with no blistering, and it should last a lifetime," says De Luca. "There are still some areas of blistering that weren't covered with the grafts, and there are other tissues like the mouth mucosa that we couldn't treat, but although we didn't completely cure the disease, we still fixed 80% of his skin."

FROM GRAFTS TO PATCHES

Over at the University of Chicago in Illinois, Xiaoyang Wu is generating genetically modified skin with a different purpose in mind. In 2017, he and his team showed that genetically modified skin grafts could be used as living 'drug patches' in mice[4], akin to plastic nicotine or hormone patches.

Using the gene-editing technique CRISPR–Cas9, the researchers modified epidermal stem cells with a version of the gene encoding GLP1 — a hormone that controls blood sugar levels and suppresses appetite — which could be switched on by the antibiotic doxycycline. They then grew the cells into small skin grafts and transplanted them onto the backs of mice.

The researchers found that the engineered skin grafts could successfully secrete GLP1 into the animals' blood in response to the drug, slowing weight gain and preventing diabetes in mice kept on a high-fat diet.

Wu's team has now used this technique to create similar patches of CRISPR-modified skin cells that produce a tweaked version of an enzyme called BChE, which breaks down cocaine[5]. Wu's version metabolizes the drug more than 4,000 times faster than the naturally occurring form, rapidly clearing it from the body and quickly killing the 'high'.

When tested in mice, the skin patch stopped the animals from becoming addicted to cocaine and prevented them from overdosing, pointing towards a potentially promising treatment for people with drug addictions. Wu and his team are also working on skin patches that could serve as long-term living biosensors — for example, engineering cells that change colour or fluoresce in response to blood glucose levels.

"Many researchers are focusing on gene therapy for internal organs like the liver, but the skin is much easier — we can culture the cells indefinitely and do the editing outside the body," Wu explains. "We can also very carefully choose the correct clones to grow up into patches, with no off-target effects or rogue genetic changes."

But human trials are likely to be some years off. "Right now we are still at the proof-of-concept stage," Wu says. "Once the technology is more established and we are confident in the procedure, we can think about moving into clinical trials to treat diseases."

Although De Luca finds this idea intriguing, he is more focused on making genetically modified skin replacement a viable treatment for the thousands of children born every year with genetic skin disorders. He is currently running two clinical trials for people with different forms of JEB, but is keen to expand into other forms of epidermolysis bullosa, which can be caused by a fault in any one of at least 18 different genes and affects around 1 in every 20,000 children born in the United States. And it's by focusing on the youngest patients, who have the most to gain from early intervention, that De Luca hopes to make the biggest difference.

"If we treat these children as soon as we can, we will prevent the formation of skin lesions rather than having to cure them — and, obviously, we need to grow less skin to cover them," he says. "If you asked me 30 years ago if it was realistic to replace the whole skin with transgenic epidermis, I would have said no, but we have done it. The final aim of my career is to make this gene therapy a real treatment for children — not a clinical trial or a demonstration of what we might do, but something that is used to treat everyone who needs it."

"After three years, his skin is stable with no blistering, and it should last a lifetime."

Three years on from his record-breaking skin replacement, Hassan is living testament to this possibility, regularly visiting the team in Modena for check-ups.

"When he was in hospital he weighed just 17 kilos and was dying, but now he is growing up," De Luca says proudly. "I last saw him two weeks ago and he is like a mascot for the institute — there is a big celebration every time we see him, and everyone who was involved in his treatment wants to give him a hug." ■

Kat Arney *is a science writer and broadcaster living near London.*

1. Hirsch, T. *et al. Nature* **551**, 327–332 (2017).
2. Krueger, G. G. *et al. J. Invest. Dermatol.* **103**, 76S–84S (1994).
3. Mavilio, F. *et al. Nature Med.* **12**, 1397–1402 (2006).
4. Yue, J., Gou, X., Li, Y., Wicksteed, B. & Wu, X. *Cell Stem Cell* **21**, 256–263.e4 (2017).
5. Li, Y. *et al. Nature Biomed. Eng.* https://doi.org/10.1038/s41551-018-0293-z (2018).



Around 80% of Hassan's skin was replaced with genetically modified skin grafts.

© 2018 Springer Nature Limited. All rights reserved.



SAM FALCONER

IMMUNOLOGY

A genetic shortcut

Gene therapies that turn the body into a designer antibody factory could bypass drawbacks of expensive treatments.

BY AMANDA KEENER

Outbreaks of infectious disease are becoming more common in many parts of the world. Between 1980 and 2010, the number of outbreaks reported worldwide more than tripled every five years. Unexpected outbreaks caused by viruses such as Ebola and Zika have led researchers to seek faster and cheaper strategies for addressing pathogenic agents they know little about. These strategies include using laboratory-made, monoclonal antibodies that can immediately bind to and neutralize specific viruses or bacteria in a person who has been infected, but also protect, for a time, anyone who is likely to be exposed to a particular pathogenic species.

But monoclonal antibodies are expensive to produce, must be stored in the cold and often require repeated administration by injection to work. That's not to mention the one to two years it takes to grow the cells that produce such antibodies and to purify and test the resulting proteins. "There's a short window of opportunity one has to halt an emerging infectious-disease breakout, and making antibodies takes time," says Neal Padte, chief operating officer at biotechnology company Renbio in New York City.

Padte belongs to a growing group of researchers who want to skip those steps by simply giving the body the genetic information it needs to make the antibodies. This can be achieved by delivering the DNA that encodes those antibodies to the cell nucleus — a process called antibody gene transfer. It's similar to the idea behind DNA vaccines, which deliver DNA that encodes vaccine components to cells. The approaches differ in that DNA vaccines are designed to trigger the immune system to make its own antibodies, whereas antibody gene transfer aims to introduce antibodies without inciting such an immune response.

Taking notes from the fields of DNA vaccines and gene therapy, researchers are working to bring treatments based on antibody gene transfer into clinical trials, using infectious diseases as a proving ground. The approach also holds promise for tackling non-infectious conditions such as cancer. "Wherever antibodies work, we believe this technology can work in the same way," Padte says.

Antibody gene transfer has to overcome the same hurdles relating to safety and delivery as does any other gene therapy, as well as more-specific challenges such as getting cells that don't normally make antibodies to produce them in large quantities. "We know it works [in mouse models]. You can do it for another thousand disease indications and it will work every time," says Kevin Hollevoet, an immunologist at the University of Leuven in Belgium. The big question, he says, is whether the approach can be applied to people.

PICK-YOUR-OWN ANTIBODIES

David Weiner, director of the Vaccine and Immunotherapy Center at the Wistar Institute in Philadelphia, Pennsylvania, has devoted almost three decades to developing and refining DNA-vaccine technology. But about eight years ago, Weiner realized that his work could make an impact in a very different field. His then-teenage daughter was diagnosed with severe Crohn's disease, and the only treatment that worked for her was a monoclonal-antibody drug that had to be injected several times a month. Weiner took notice of the fast growth of therapies based on monoclonal antibodies, which include anti-inflammatory drugs such as adalimumab (Humira) and checkpoint inhibitors such as pembrolizumab (Keytruda). "It's one of the most important fields in biotech," Weiner says.

The drugs that the field produces are also among the most expensive. Costing up to US$100,000 per year of treatment, monoclonal-antibody therapies are out of reach for most of the world's population. Weiner thinks that gene therapy could make such drugs more accessible. It costs much less to make DNA in the lab than to produce monoclonal antibodies. The approach would also require fewer doses because lab-made DNA can last for weeks to months in the cell nucleus, while continuously instructing the cell to churn out antibodies.

Since 2013, Inovio Pharmaceuticals in Plymouth Meeting, Pennsylvania, a company co-founded by Weiner, together with Weiner and his team at Wistar, has been developing a number of DNA-encoded monoclonal antibodies. It started by creating antibodies to tackle viral infectious diseases such as chikungunya and dengue fever and has now broadened its scope to develop such antibodies against antibiotic-resistant pneumonia and two proteins found at elevated levels in tumours of the prostate gland. They are now working on DNA-encoded monoclonal antibodies that mimic antibodies against the Ebola virus from

© 2018 Springer Nature Limited. All rights reserved.

CAROLINE SINNO PHOTOGRAPHY

the blood of people who survived infection.

Inovio is not alone. Several groups of researchers have produced monoclonal antibodies in mice that can protect the animals from infection and attack tumours. For example, Padte and his collaborators in the United States and China delivered genes that encode the three antibodies that comprise the anti-Ebola-virus vaccine ZMapp, as well as three anti-influenza antibodies, into mice[1]. The antibodies protected the animals from both Ebola and influenza.

This ability to pick and choose the most effective antibodies for a disease is especially attractive to researchers who study a special class of antibody that can neutralize multiple strains of HIV. Up to one-third of people with the virus make these antibodies. That could be down to genetic differences between individuals; it might also relate to the strain of HIV encountered. "What you can do with antibody gene transfer is just take the successful antibodies that came out of these unusual pairings of people and viruses and give them to a broad audience," says Alejandro Balazs, who studies immunity against HIV at the Ragon Institute of MGH, MIT and Harvard in Cambridge, Massachusetts. That way, he says, "You are taking the black box of the immune response out of the equation."

The US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, is testing the delivery of a gene that encodes one such neutralizing antibody, on which Balazs has worked for more than ten years. The trial will evaluate the therapy's safety in people with HIV. If it goes well, Balazs says, there might be opportunities to check whether the participants' bodies are converting the gene into the desired antibody. A separate trial, run by the International AIDS Vaccine Initiative in New York City, is testing the safety of another gene encoding an HIV-neutralizing antibody in a cohort of healthy men. The outcomes of both HIV trials will signpost how well antibody gene transfer works in humans. "A lot of people are looking at this very closely," Balazs says.

IT'S ALL IN THE DELIVERY

There are many ways of delivering genes to cells. Few have been tested in people, however, and none has been assessed for inducing antibody production. The HIV trials use a virus called adeno-associated virus (AAV) to carry genes encoding HIV-targeting antibodies into the muscle cells of participants. AAV has a knack for getting foreign DNA into human cells, says Ronald Crystal, who works on gene therapy at Weill Cornell Medicine in New York City. "That's what they live for."

AAV is also well suited to inserting antibody genes into hard-to-reach organs such as the brain. Crystal and his collaborators used the AAV approach to deliver an antibody that reduced levels of tau, a protein implicated in Alzheimer's disease, into the brains of mice with another type of dementia[2].

But AAV, as well as other viruses used in antibody gene transfer, has downsides. It can incite an immune response. And because the virus is grown inside cells, production can be time-consuming and costly. Approaches that leave out viruses, such as Weiner's DNA-encoded monoclonal antibodies, avoid those limitations. But without a virus to transfer the DNA, cells have to be coaxed into accepting foreign genes, usually by a process called electroporation, in which an electric current is used to create tiny, temporary holes in cells through which DNA can pass.



Neal Padte (left) is building DNA constructs that could enable the body to produce tailored antibodies.

Scancell, a cancer-immunotherapy company in Oxford, UK, has used electroporation to transfer a gene encoding a lab-designed antibody that primes immune cells called T cells to target tumours in people with melanoma. In 2017, the company reported that the treatment safely induced an immune response against the cancer.

For an even simpler approach to delivering antibody genes, others are turning to messenger RNA — the molecule that conveys information stored in DNA to the cellular machinery that makes proteins. For reasons not fully understood, mRNA can make its way into muscle cells without the need for electroporation.

In 2017, Drew Weissman at the University of Pennsylvania in Philadelphia and his collaborators injected an mRNA sequence for an HIV-neutralizing antibody into mice, protecting the animals from infection with HIV[3]. The biopharmaceutical company CureVac in Tübingen, Germany, and its collaborators reported success with mRNA-encoded antibodies against viral proteins involved in influenza and rabies, as well as the mRNA-encoded monoclonal-antibody drug rituximab, which is used to treat non-Hodgkin's lymphoma[4]. And BioNTech in Mainz, Germany, is experimenting with mRNA as a means of introducing T-cell activating antibodies for cancer immunotherapy[5].

THE HUMAN PROBLEM

As antibody gene transfer enters clinical testing for infectious diseases and cancer, some researchers are starting to consider how to make it work for chronic conditions such as arthritis. This is more challenging because people with such disorders often have to switch between monoclonal antibodies to find the one that works best. A therapy that enables the body to produce antibodies for up to years at a time, as can be the case with AAV-delivered genes, would remove that option. "There is the risk that you can't shut it off," says Crystal.

Balazs and other researchers are working on 'off switches' in the form of complementary gene therapies or drugs. But for now, Balazs says, it is still unclear whether approaches that have been successful in mice will work in humans. "We're asking this one site of muscle to pump out enough antibodies to distribute to the entire body," says Hollevoet, who is studying sheep to get a better sense of how much antibody the human body might produce.

For the antibodies that have been tested only in animals, it's impossible to know the concentration in blood that will be needed to treat a given disease. "That's why these first clinical trials are going to be so important," says Balazs. Then researchers can deal with next-level features such as off switches. The mission is straightforward, he says: "Let's just see if we can make the thing turn on." ■

Amanda Keener *is a freelance science writer in Littleton, Colorado.*

1. Andrews, C. D. *et al. Mol. Ther. Methods Clin. Dev.* **7**, 74–82 (2017).
2. Liu, W. *et al. J. Neurosci.* **36**, 12425–12435 (2016).
3. Pardi, N. *et al. Nature Commun.* **8**, 14630 (2017).
4. Thran, M. *et al. EMBO Mol. Med.* **9**, 1434–1447 (2017).
5. Stadler, C. R. *et al. Nature Med.* **23**, 815–817 (2017).

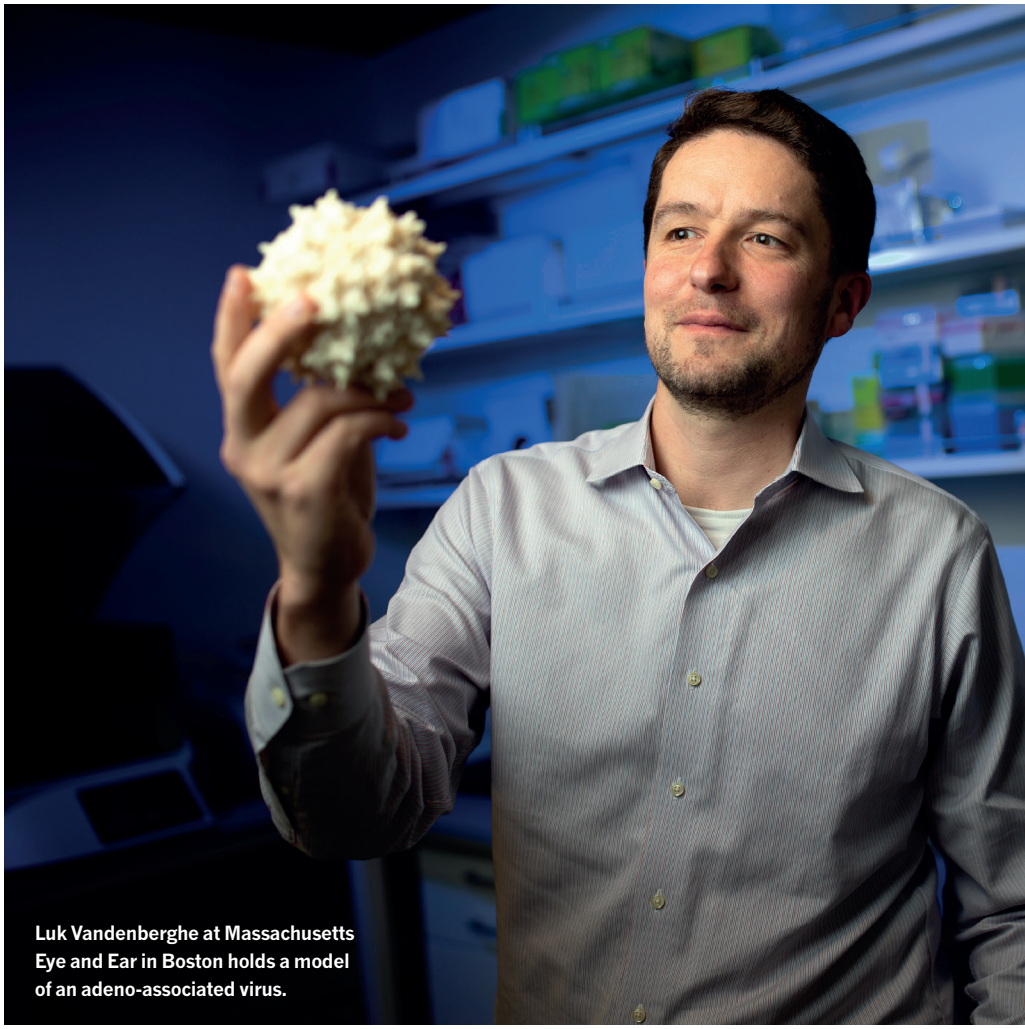
© 2018 Springer Nature Limited. All rights reserved.

THERAPEUTICS

Special delivery

By tweaking a virus's shell, Luk Vandenberghe thinks he can transport genes into cells much more efficiently and cost-effectively.

BY NEIL SAVAGE



Luk Vandenberghe at Massachusetts Eye and Ear in Boston holds a model of an adeno-associated virus.

ARAM BOGHOSIAN/MASSACHUSETTS EYE AND EAR

Luk Vandenberghe walks over to a shelf in his office and picks up two fist-sized objects. One is a more complicated version of a Rubik's Cube, with 20 individually coloured sides instead of the standard 6. The other is an off-white glob of hard plastic produced by a 3D printer. It's studded with bumps, dimples and repeating triads of vaguely pyramid-like shapes, 20 in all.

Both are models of an adeno-associated virus (AAV), a favourite vector among clinicians for delivering genes to cells. Vandenberghe, a bioengineer who directs the Grousbeck Gene Therapy Center at Massachusetts Eye and Ear in Boston, is trying to work out what effect all those tiny structures have on the behaviour of the virus. His aim is to manipulate them to improve the vector's ability to deliver genes without, in essence, messing up the colour pattern on the Rubik's Cube — or in this case, the icosahedron.

Vandenberghe completed his doctorate on the structural basis of AAVs in 2007 at the Catholic University of Leuven in Belgium, and later went on to become an associate professor at Harvard University in Cambridge, Massachusetts. Through a mix of computational modelling and DNA synthesis, he has been trying to solve the problems that arise from using natural AAVs for gene therapy, and has founded three companies to bring his technologies to market. One of them is using an unusual non-profit approach to tackle the economics of developing gene therapy for extremely rare diseases.

Naturally occurring AAVs have become a workhorse of gene therapy. They infect human cells without causing illness, and different variations of the virus target different cell types — so selecting the right virus is essential for getting replacement genes to cells where they are needed. Vandenberghe and his colleagues have so far identified more than 140 natural variations of the virus[1].

But scientists would like to fine-tune AAVs to improve their specificity and the efficiency with which they penetrate tissue. The goal of AAV research over the past two decades has been treatments that use lower doses and do not affect off-target tissues.

Researchers are also trying to solve another problem. Because the viruses circulate in the wild, many people have been exposed to them and have developed immunity. That puts therapies that rely on AAVs out of reach for many patients. Estimates for the number of people with immunity vary widely, Vandenberghe says, from 20–90%. Some of that variation is due to geography; the viruses are more prevalent in Africa, for instance, than in the United States.

Bioengineers think they can achieve large changes in the function of AAVs by altering the capsid — the protein shell of the virus. For instance, capsid differences are the reason why one naturally occurring AAV targets liver cells with up to 100 times the efficiency of another. "Unfortunately, we still don't know exactly what it is that makes one virus go to the liver 100-fold better than the other," Vandenberghe says. Scientists also don't fully understand how a change in one part of the virus might affect the structure in another part, in much the same way that moving a red square on a Rubik's Cube might put a green square on another face out of place. "What we're trying to do is exactly solve that Rubik's Cube dilemma," says Vandenberghe. "That's not trivial on a cube, and it is certainly not trivial on an icosahedron."

LEARNING FROM HISTORY

To learn more about how structure affects function, Vandenberghe and his team decided to reconstruct the evolutionary history of AAVs. In 2015, he and his colleagues fed the protein sequences of 75 AAV variants isolated from human and non-human primate tissues into an evolutionary computer simulation and reconstructed the sequences of nine possible ancestors of modern AAVs[2], the oldest of which they named Anc80. Vandenberghe is not claiming these are the actual forms of previous generations of viruses, but that isn't the point, he says. "We didn't quite care. What we really wanted to do was find inroads into this structural problem that we had."

On the basis of the sequences, the researchers synthesized the ancestral viruses and examined their characteristics — and Anc80 proved to be especially interesting. When injected into mice, the virus was able to penetrate all of the hair cells in the inner ear and most of the hair cells in the outer ear, something no previous virus had accomplished. In 2017, Vandenberghe and his colleagues used Anc80 in mice to treat a genetic

© 2018 Springer Nature Limited. All rights reserved.



disorder called Usher syndrome that causes deafness and visual impairment[3]. Excited by the potential of such a vector, Vandenberghe and his colleagues founded a company, Akouos, in Boston to develop treatments for hearing loss. In August, the start-up secured US$50 million in a first round of investment.

Vandenberghe's team is also collaborating with Selecta Biosciences in Watertown, Massachusetts, which wants to develop gene therapies using Anc80. Vivet Therapeutics in Paris is licensing the vector for use in developing treatments for inherited liver disease. And Lonza in Basel, Switzerland, is licensing the technique for making the virus so it can manufacture the vector for drug-makers. Back in 2011, before the Anc80 work, Vandenberghe also co-founded GenSight Biologics in Paris to develop treatments for rare inherited retinal diseases; the company currently has two drugs in clinical trials.

Creating better vectors is the key to expanding gene therapy, says Eric Kelsic, a systems biologist in the laboratory of molecular engineer George Church at Harvard University. Kelsic is taking a data-driven approach to capsid engineering. He selects an amino acid from the protein sequence of an AAV and systematically switches it with each of the other 19 amino acids in existence in turn to see what changes. Then he moves on to the next amino acid in the sequence and repeats the process. "With this approach, we know what the effect is for every possible individual change," he says. Using machine learning, he predicts what will happen when single-amino-acid changes are combined, then synthesizes promising sequences and tests the AAVs in mice or non-human primates.

Kelsic and Church have founded a company, Dyno Therapeutics in Cambridge, Massachusetts, to create vectors this way. Kelsic predicts that even for tissues such as the brain that can already be targeted with AAVs, more-efficient viruses will lead to improved therapies. The greater achievement, however, will be the ability to target organs that are currently hard to treat, such as the lung and kidney. "As we improve delivery further it will enable new therapies which just aren't possible today," he says.

A DIFFERENT BUSINESS MODEL

The companies that these researchers have founded follow the standard for-profit model used by most biotechnology start-ups. But Vandenberghe is taking a different approach with Odylia Therapeutics, a not-for-profit company he founded in February. Odylia aims to develop therapies for what Vandenberghe calls "ultra-rare" genetic causes of blindness, which he defines as those that affect 3,000 or fewer people in the United States. The firm is supported financially by Massachusetts Eye and Ear and the Usher 2020 Foundation in Atlanta, Georgia, a charity focused on curing the sight loss caused by Usher syndrome. One of the charity's founders, Scott Dorfman, who has two children with Usher syndrome, is chief executive of Odylia.

So far there is only one available gene therapy for blindness. In late 2017, the US Food and Drug Administration (FDA) approved voretigene neparvovec (Luxturna) for the treatment of eye disease caused by a mutation in the *RPE65* gene, which normally produces a protein in the thin layer of cells at the back of the eye. As a proof of concept, the treatment shows that gene therapy can be used to cure eye disease. But mutations in more than 200 genes have been linked to hereditary eye diseases, and Vandenberghe says that there is little appetite in the pharmaceutical industry for developing individual therapies to correct many of the other genes.

"As we improve delivery further it will enable new therapies which just aren't possible today."

It can cost millions of dollars to develop a drug and take it through clinical trials, and if a disease is rare, it may not make economic sense for companies to pursue a treatment for it. That is a particular issue in gene therapy, in which people are often cured with a single dose rather than a life-long drug regimen. The doses required for eye diseases are tiny because the retina is a relatively small organ, and some retinal diseases are so rare that it's possible that a single batch of the drug could treat the entire patient population in the United States, Vandenberghe says.

A WIDER CONCERN

The question of how to develop gene therapies for rare diseases is of great concern to the US National Institutes of Health, says P. J. Brooks, program director at the institute's Office of Rare Diseases Research in Bethesda, Maryland. "When people discuss business models around treatments for rare diseases, the basic assumption is that there is a business model," he says. "But for some of these diseases where there's a very small patient population, there may not be one." Brooks says Odylia is the first company he has heard of to try this non-profit approach.

The idea, Vandenberghe says, is to find economies of scale by sharing resources and scientific and commercial expertise across the development of a range of drugs that are similar to one another. If the same group of people develops the drugs, designs the clinical trials and produces the materials, there should be less duplication of effort, he notes. Vandenberghe also hopes that after creating two or three successful treatments, the company will be able to provide data to convince the FDA that there are enough similarities between the drugs to enable them to use experience with one drug to help establish the safety and efficacy of another. It is also possible that Odylia will take development of a drug far enough in this model that a for-profit company will decide to buy it and complete the work, providing funding for Odylia while reducing the pharmaceutical company's costs and risks.

If Odylia does bring a drug to market, it will probably be sold at cost, Vandenberghe says. That could still be expensive, but possibly less so than if it had been developed the usual way. There is also a chance that if a drug candidate gets through phase I and II clinical trials, the FDA could allow it to be provided on a compassionate-use basis without a final clinical trial, or that most patients could be treated as part of an open-ended trial.

If the model is successful, it could be extended to other rare, single-gene disorders and perhaps provide insights for developing gene therapies for more common conditions. "Maybe this is one of those areas where industry can acknowledge that this is indeed non-competitive," Vandenberghe says. Ideally, he says, that would set up a happy scenario. "We can all come together around some of these common goals, apply them to ultra-rare diseases, and then take those lessons to the more commercial world afterwards." ■

Neil Savage *is a science and technology journalist in Lowell, Massachusetts.*

1. Gao, G. *et al. J. Virol.* **78**, 6381–6388 (2004).
2. Zinn, E. *et al. Cell Rep.* **12**, 1056–1068 (2015).
3. Pan, B. *et al. Nature Biotech.* **35**, 264–272 (2017).

© 2018 Springer Nature Limited. All rights reserved.

REGULATING A REVOLUTION

Health authorities wade into the flood of gene therapies.

BY ERIC BENDER



SAM FALCONER

© 2018 Springer Nature Limited. All rights reserved.

For rare genetic diseases that affect the young, such as a neurodegenerative condition called spinal muscular atrophy, gene therapies bring much-needed hope — a chance for the child to live a relatively normal life. But they also raise serious fears about their efficacy and the potential risks that accompany irreversible one-off treatments.

The responsibility for balancing these hopes and fears lies with the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA). Their credentials as gatekeepers to the therapies will soon be tested by a flood of clinical trials. This year the FDA expects to receive about 250 applications to start clinical trials for novel cell and gene therapies, says FDA commissioner Scott Gottlieb.

Faced with rapid advances in biological understanding and therapeutic delivery technologies, the two regulatory agencies are establishing new guidelines for clinical trials and are preparing to make tough decisions about which drugs to approve for marketing. But drawing on their experience with hundreds of earlier studies, the agencies are confident that they can assess gene therapies as effectively as they do any other novel therapeutics.

STANDARDIZING SAFETY

Gene therapy has long been haunted by a very small number of deaths, originally in a 1999 US clinical trial and then in a European study a few years later. However, a series of successful clinical trials over the past decade has created sufficient confidence to move forward with these treatments.

One milestone, in December 2017, was the first FDA approval of an *in vivo* gene-therapy product, for Luxturna from Spark Therapeutics, based in Philadelphia, Pennsylvania. Luxturna treats a rare, inherited eye condition caused by mutations to a gene called *RPE65* that can cause blindness.

Another was the announcement in August 2018 that gene therapies no longer need to be reviewed before clinical studies can begin by a US National Institutes of Health (NIH) advisory committee on recombinant DNA that was created at the dawn of genetic medicine. "There is no longer sufficient evidence to claim that the risks of gene therapy are entirely unique and unpredictable — or that the field still requires special oversight that falls outside our existing framework for ensuring safety," wrote Gottlieb and NIH director Francis Collins in a paper published earlier this year (F. S. Collins & S. Gottlieb *N. Engl. J. Med.* **379**, 1393–1395; 2018).

Even so, such a new class of medicines still poses serious risks. "It's not that people say: 'Oh, it's all safe, don't worry'," says Katherine High, a haematologist and president of Spark. "It's that now we really have some parameters inside which we can work."

She points out, for example, that previous trials have gathered plenty of evidence about therapies such as Luxturna that are delivered by adeno-associated viruses (AAV), especially for systemic administration or for commonly targeted tissues such as the eye. Such AAV therapies often create a short-term immune response in the liver, but this problem can generally be treated by using steroids. "For other target tissues, or for doses that are higher than people have used to date, you may need additional information," High says. "There actually are a wealth of approaches to overcome immune response, and it's a matter of doing the clinical investigations and finding answers."

Barry Byrne, director of the Powell Gene Therapy Center at the University of Florida in Gainesville, says it is far too soon to declare today's gene therapies safe. "There's very limited experience," he cautions, "and there's much more work to be done to understand how these might be used in a variety of conditions."

There are many unanswered questions, such as what happens if a patient who receives a gene therapy delivered by AAV has previously been exposed to some form of the virus, or if proteins created by gene therapies provoke a reaction because the immune system has not been trained to recognize them as 'self', Byrne adds. But he believes that strategies are emerging to avoid or control such immune problems.

> "YOU CANNOT HAVE TWO STANDARDS FOR SAFETY."

New forms of gene-therapy delivery and mechanisms of action sometimes do not perform as expected when they enter clinical studies. In September 2018, Sangamo Therapeutics, based in Richmond, California, reported the initial results of the first trial of gene editing inside the body, for a therapy to treat a rare metabolic disease called Hunter syndrome. The disease, which primarily affects males, causes a host of serious symptoms, and treatment currently requires weekly injections of enzymes. But the initial Sangamo trials failed to demonstrate clinical benefit, and they are now continuing with higher doses.

The regulatory agencies are seeking to provide more guidance on such emerging gene-editing therapies. The EMA and the FDA are working together "to avoid digressions between the two of us", says Hans-Georg Eichler, senior medical officer at the EMA. "In gene therapy in general, we like to believe that we know what the major risks are, but you can never know," Eichler says. "Tomorrow, something totally new could come out of the blue. But that doesn't say that gene therapy shouldn't be made available to patients."

BETTER BY DESIGN

Given the novelty and the potential risks and rewards of gene therapies, their sponsors tend to start working with regulatory agencies early in development — often, very early. "Ideally, you talk with the agencies when you are designing your preclinical development," says Anne-Virginie Eggimann, vice-president for regulation at biotech company Bluebird Bio in Cambridge, Massachusetts. "You can have a general discussion with them on designing that programme, as well as how you see your first-in-human clinical trial." In October, Bluebird Bio submitted a marketing application to the EMA for its LentiGlobin gene therapy, which is designed to treat a rare blood disease called transfusion-dependent β-thalassaemia.

Like LentiGlobin, about 70% of the investigational new drug (IND) applications for gene therapy submitted to the FDA are for rare diseases. Most of these conditions first appear in childhood, and most of those have devastating results. But running a normal clinical trial, which includes large numbers of subjects and a control arm, is often impossible.

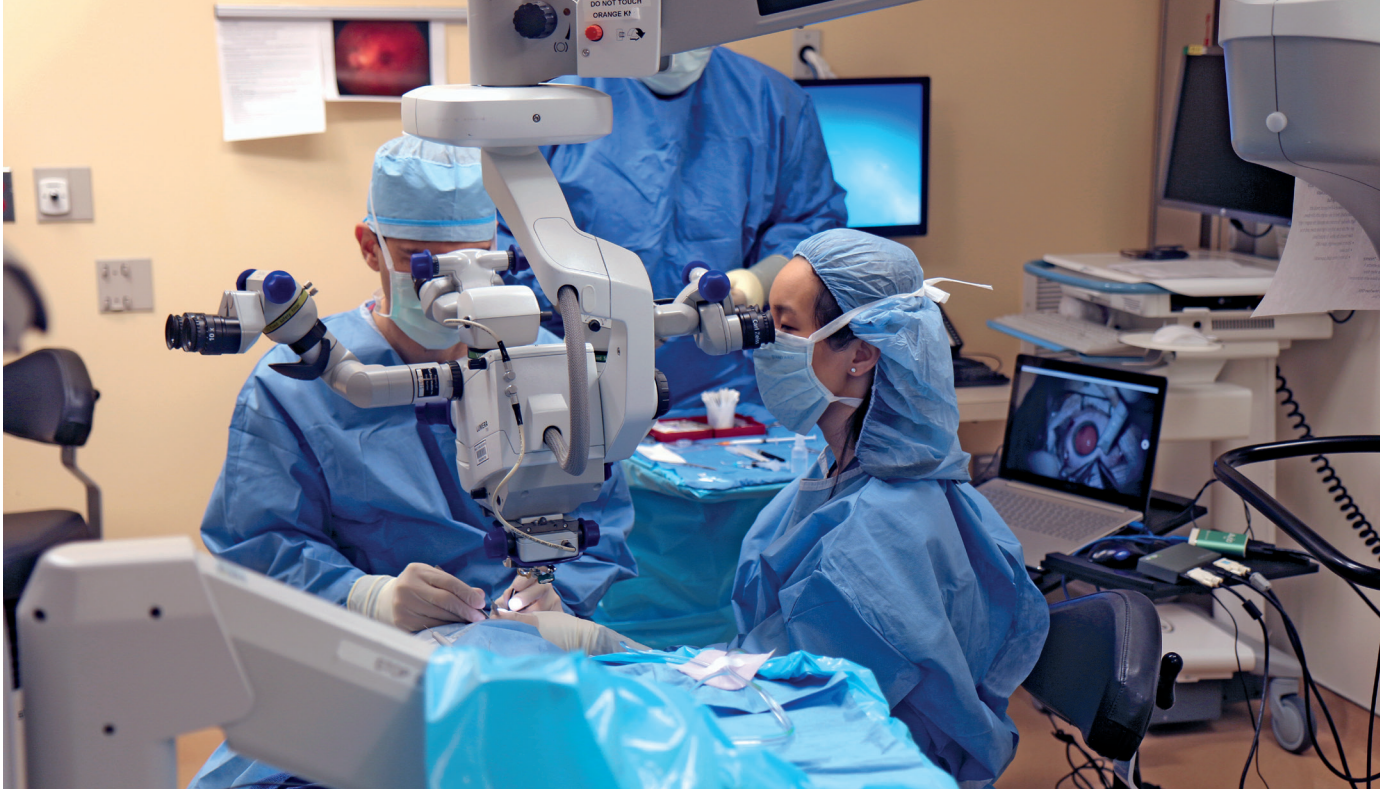
"We know that in these situations you have to exercise some flexibility, and that is exactly what we usually discuss with the companies when they come early," says Eichler. "We negotiate and see how can we get the best that is doable in the circumstances."

Given the devastating nature of many rare inherited diseases that strike children, parents often press for accelerated clinical tests. But developers emphasize that lowering safety standards is not an option. "I really understand the urgency of parents whose child has a serious illness," says High. "On the other hand, this is a field where you cannot have two standards for safety."

Trial sponsors and regulatory agencies also worry about how candidate products are manufactured, and how the products might be affected by changes in the manufacturing process over time. Making gene therapies is a highly complex process using biological materials, and extremely high quality must be assured at every step. Most academic labs and biotech startups lack the expertise and the equipment to pull off this feat well enough to produce commercial-grade therapies at a commercial scale. Few biomanufacturing facilities currently provide such services, and these operations are overloaded by the number of therapies now heading towards clinical trials. The difficulties are compounded by the need, as trials progress, to improve the manufacturing processes while keeping the product consistent enough to keep regulators happy.

"Manufacturing is something we will have to think about differently, so we can get it right the first time," says Peter Marks, director of the FDA's Center for Biologics Evaluation and

© 2018 Springer Nature Limited. All rights reserved.



ED SHIPMAN/MASSACHUSETTS EYE AND EAR

Surgeons use Luxturna, the first *in vivo* gene therapy to be approved by the US Food and Drug Administration, to treat a boy with a genetic eye condition.

Research in Silver Spring, Maryland, which oversees gene therapies.

"Quite often people develop things on the lab bench at a very small scale, and they need to scale up and scale out their thinking," says Jacqueline Barry, chief clinical officer for the Cell and Gene Therapy Catapult, a UK government commercial incubator. "We try to work with them very early on about moving to a good manufacturing process and gathering data that will support the evolution of the product between clinical-trial phases without having to go back and redo studies."

Gene therapies also require follow-up for patients that extends for years after product approval because the long-term effects of these one-time treatments are simply not known. "Clinicians must come to grips with that idea," says Eichler. "As we treat, we must ascertain that the patient experience — good or bad — must somehow be fed back to decision-makers and contribute to long-term knowledge generation."

SEEKING APPROVAL

Europe and the United States have very different legal and regulatory regimes for approving gene therapies. The main difference is that the FDA oversees clinical trials, whereas the EMA does not. To run a clinical trial in any of the 28 members of the European Union, "you have to get approval from a competent authority and from the ethics committee in that member state," says Barry. You also have to get approval for using a genetically modified organism (GMO). However, "the clinical-trial directive and the GMO directive are translated slightly differently in each country," she points out.

Moreover, participation in decisions is structured differently in Europe and the United States, says Eggimann. At the EMA, committee members from various states meet to make decisions about marketing approval. At the FDA, reviewers within the appropriate division follow the drug candidate throughout its entire life cycle.

But the two agencies take similar data-driven approaches to assessing drug safety and efficacy, often actively working together in the process. Several times a year, for example, they hold teleconferences on gene therapies. "We all know there are so many uncertainties in this field, and so many new developments that we want to keep each other abreast of," says Eichler.

> "I DON'T SEE THE AGENCIES AS A BARRIER AT ALL."

Both agencies released major updates to their gene-therapy guidelines in 2018. The FDA, for example, offered its first draft recommendations by class of illness, starting with haemophilia, retinal disorders and rare diseases. It also added draft frameworks for certain manufacturing processes and requirements for long-term patient follow-up. The EMA also completely overhauled its frameworks for gene therapies. For instance, it reworked its guidance on the design, manufacture, characterization and testing of delivery mechanisms.

"As the field gains more and more experience, the broad outlines of what needs to be submitted to initiate clinical studies have come more clearly into focus," says High. "You find that reflected in the guidance documents that the FDA and the EMA provide."

Gene-therapy developers worry that the agencies lack enough experts to deal with the incoming wave of trials for cell and gene therapies, which the FDA estimates will reach 1,000 a year by 2021. "They don't have enough people to handle that kind of workload," says High.

"For the FDA, the issue is always around the budget, and being able to have the appropriate technology and people to deliver on their commitments," says Peter Saltonstall, president of the National Organization for Rare Disorders based in Danbury, Connecticut.

It is still early days for gene therapies, but so far, developers generally give both agencies high marks as partners. "I don't see the agencies as a barrier at all," says Byrne. "They have so many mechanisms for interacting with sponsors now, and they've always approached sponsors as collaborators in bringing these agents forward."

Eggimann agrees. "The regulators have been very supportive of innovation and gene therapy in general, and they are very eager to learn," she says. "Our challenge comes from the novelty of the science, not so much from the regulatory aspects."

Meanwhile, the therapies keep moving forward. Among them is AVXS-101, a gene therapy from AveXis based in Bannockburn, Illinois. AVXS-101 has raised high hopes in early clinical trials for the treatment of spinal muscular atrophy, that devastating neurodegenerative condition that affects children. In October 2018, AveXis applied to both the FDA and the EMA for marketing approval — yet another bridge that gene therapy is crossing on its journey from the lab to the clinic. ■

Eric Bender *is a science journalist based in Newton, Massachusetts.*

© 2018 Springer Nature Limited. All rights reserved.

PERSPECTIVE



Access and affordability for all

The hope of gene therapy could be crushed by its financial burden unless there are more rational ways of paying for it, says **Michael Sherman**.

Gene therapy offers the possibility of a cure for previously untreatable diseases. But although the science and technology behind it are awe-inspiring, the costs can be daunting. Treatments are likely to have a price tag in the neighbourhood of US$1 million or more — a cost that is ultimately borne by all individuals, not just patients, through taxes and insurance premiums.

In the United States, which lacks government-administered provision of universal health care, there is a strong expectation that health insurers will pay for therapies that have been approved by the US Food and Drug Administration (FDA), particularly if a treatment is the only effective one for a given malady. In cases in which the efficacy data and value proposition are questionable, FDA approval can create enormous pressure to provide coverage.

Some stakeholders — including pharmaceutical companies and government policymakers — have been squeamish about introducing measures of cost effectiveness into the decision-making process because of concerns that such an approach could lead to putting a price on life and, ultimately, the rationing of care. Unfortunately, this has had an unintended consequence: it has led to a system that has no mechanism for imposing price ceilings. Many individuals in the United States see substantial cost increases for their medications year after year.

One possibility would be for the FDA to consider a pathway in which it expedites approval for a treatment in the absence of sufficient high-quality data, particularly for rare diseases that have no effective treatment, in return for the drug maker agreeing to a so-called value-based agreement that would tie reimbursement to the success of the drug. When treatment works, the manufacturer would receive full payment. When the patient shows a limited response to treatment, there would be a partial payment. And when the treatment fails altogether, no payment would be made.

I work for the health insurer Harvard Pilgrim Health Care in Wellesley, Massachusetts, and in January my company entered into a value-based agreement with Spark Therapeutics in Philadelphia, Pennsylvania, for the gene therapy voretigene neparvovec (Luxturna), a treatment for a form of hereditary blindness. This agreement is already driving considerable discussion between payers and pharmaceutical companies that have upcoming gene therapies and other high-cost, innovative treatments. Other firms have forged similar deals. For example, in 2016, the pharmaceutical company Novartis in Basel, Switzerland, signed a deal with several insurers, including Cigna in Bloomfield, Connecticut, and Harvard Pilgrim, for its combination drug sacubitril–valsartan, a treatment for heart failure. In the event that people receiving the drug fail to show a reduced rate of hospitalization for heart failure in clinical trials, the drug cost will be reduced. Collaborative deals such as this give hope that stakeholders will work together to ensure that all who might benefit have access to cutting-edge medical advances.

Gene therapy, which offers the potential of extremely effective but extremely expensive treatments, is a good candidate for value-based agreements. Take, for example, the high-cost biological drug eteplirsen, which targets the gene responsible for Duchenne muscular dystrophy (DMD). The FDA expedited approval of the drug in 2016 because DMD was a fatal, progressive disease with insufficient treatment options. Approval was granted despite the FDA's advisory committee voting against it and despite slim evidence of efficacy — the pivotal trial, which enrolled just 12 boys, showed very small changes in the surrogate measure used as an outcome.

The agency's decision sent shock waves through the US insurance industry and led to variability in coverage policies. Many companies agreed to pay for the drug, which costs around $300,000 per year, but others initially declined to do so.

In this case, a value-based agreement could have set out a multi-year payment model that would terminate if the effectiveness of the drug failed to persist over the long term. And because such a deal would enable broad access to the therapy, it would in turn generate robust real-world evidence of the treatment's efficacy. Such data could then be used to gain conventional FDA approval. Sarepta Therapeutics in Cambridge, Massachusetts, the company that developed eteplirsen, chose not to enter into value-based agreements for that drug, but it is collaborating with a partner to develop a one-time DMD gene therapy that is expected to be much more expensive. That therapy might present an opportunity to enter into an innovative financing agreement to promote access.

TREATMENTS ARE LIKELY TO HAVE A PRICE TAG IN THE NEIGHBOURHOOD OF US$1 MILLION OR MORE.

Some pharmaceutical companies oppose value-based pricing, questioning whether the approach maximizes shareholder value. It is fair to acknowledge that any solution to improve access to health-care advances should provide a reasonable return to the companies that develop such innovations. It is also appropriate to ask whether treatments for rare conditions should be priced higher to ensure that companies will pursue the development of drugs that will always have a limited market.

Whether or not we choose to acknowledge it, there is a limit to the portion of a country's gross domestic product that can be spent on health care. To balance access and affordability over the long term and ensure that our loved ones can receive the next generation of innovative therapies, payers, pharmaceutical companies and regulatory agencies need to collaborate in a way that benefits all stakeholders. Value-based agreements from the past few years provide a model that could be applied to upcoming gene therapies and other high-cost, innovative treatments. A spirit of collaboration among industry players could ensure that everyone who needs an innovative, expensive treatment can have access to it. ■

Michael Sherman *is senior vice president and chief medical officer at Harvard Pilgrim Health Care in Wellesley, Massachusetts, and a faculty member at Harvard Medical School in Boston, Massachusetts.*
e-mail: michael_sherman@harvardpilgrim.org

© 2018 Springer Nature Limited. All rights reserved.

DHIRAJ SINGH/BLOOMBERG VIA GETTY IMAGES



Crop researchers at an institute near Hyderabad, India, are using biotechnology to boost yields of pigeon-pea plants, a food crop and important protein source.

How Indian biotech is driving innovation

Bolstered by government support, a wealth of investment and an eager graduate workforce, the country's biotechnology industry is booming.

BY BIANCA NOGRADY

Anu Acharya was in her twenties when the human genome was first mapped in its entirety. In 2000, the young Indian entrepreneur was just breaking into the biotechnology arena with her first start-up — the genomics and bioinformatics company Ocimum Biosolutions in Hyderabad. She saw the Human Genome Project's achievements as opening up a new world of possibilities in personalized medicine, informed by an individual's genetic profile and predispositions — but at the time, the field of genomic medicine was dominated by Western science.

"I wanted to make sure that India had its own voice heard in that," Acharya says. So, a decade later, she launched her second biotech start-up — molecular-diagnostics company Mapmygenome, also in Hyderabad — to bring the personalized-medicine revolution to India's diverse population.

"Because, ultimately, when you're making medicine precise, it has to be for specific individuals and populations rather than based on one population that has been studied."

Acharya is among India's rapidly growing ranks of biotechnology entrepreneurs and start-ups that are riding a wave of government enthusiasm, free-flowing venture capital and growing demand from an increasingly wealthy population that wants better treatment options. These factors are helping to drive India's biotechnology industry beyond its historical focus on unbranded generic drugs and into the innovation limelight.

By the end of 2016, there were more than 1,000 biotechnology start-ups in India, and more than half of these had been established within the previous 5 years. Australia, by contrast, has 470 biotechnology companies and the United Kingdom 3,835. The biotechnology industry in India was valued at US$11 billion in 2016, and is forecast to grow to $100 billion by 2025.

More than half of the biotechnology start-ups are in the medical arena — diagnostics, drugs and medical devices — but 14% are in agricultural biotechnology, 3% in bioindustry, 1% in bioinformatics and 18% in biotechnology services.

India is already eyeing the prospect of its first biotechnology 'unicorn' — a ▶

© 2018 Springer Nature Limited. All rights reserved.



A banana farm in Jammu, northern India. A research programme used biotechnology to develop trees that would grow in the region.

SHILPA THAKUR/PACIFIC PRESS VIA ZUMA WIRE/SHUTTERSTOCK

start-up valued at more than $1 billion. The potential unicorn in question, Biocon in Bangalore, started in 1978 as an enzyme manufacturer but is now making a name for itself in the research and development of biological drugs for treating diabetes, cancer and autoimmune diseases. By March 2018, its revenue had topped $650 million.

India has long been a global player in the manufacture of generics (unbranded versions of existing pharmaceutical products), accounting for 20% of global exports of generics and earning just over $17 billion from that market in 2017. So what has prompted the nation to move beyond such a lucrative comfort zone and into the more risky game of biotechnology innovation?

GOVERNMENT SUPPORT

In 1986, with the encouragement of then-prime minister Rajiv Gandhi, India became one of the first countries in the world to have a government unit dedicated solely to biotechnology. The Department of Biotechnology started with a relatively modest budget of between 40 million and 60 million rupees ($557,000–835,000), growing exponentially to 24.1 billion rupees in 2018. In addition to establishing 17 Centres of Excellence in Biotechnology at institutes and universities around the country, the department has supported the creation of 8 biotechnology parks, or incubators, in cities such as Lucknow, Bangalore, Hyderabad, Chennai and Kerala.

The aim of these parks is to provide facilities for scientists and small to medium-sized enterprises (SMEs), where they can develop and demonstrate their technologies and even build pilot plants. The hope is that this will speed up the commercialization process. The park staff also provide mentorship and guidance on issues such as intellectual property, business plans, proposals for clinical development and exit strategies.

This support is helping to address some of the logistical challenges that have hampered industry in the past, says Tej Singh, a biophysicist at the All India Institute of Medical Sciences in New Delhi and president of the Biotech Research Society, India.

"They created some sort of industrial regions in many areas, but there were issues like electricity, water [supply]; all these small things used to take time," Singh says. "But the government has addressed these things nowadays; this current government particularly is very proactive."

The Department of Biotechnology has also supported biotechnology research infrastructure, including a high-resolution mass spectrometry facility in Mumbai, flow-cytometry, imaging and microarray facilities in Delhi, and animal-house facilities in five other regions.

The jewel in the departmental crown, and the scheme that attracts the most attention, is the Biotechnology Industry Research Assistance Council (BIRAC). This is a not-for-profit, public-sector enterprise that was set up by the Department of Biotechnology in 2012 to "stimulate, foster and enhance the strategic research and innovation capabilities of the Indian biotech industry, particularly start-ups and SMEs".

"The idea of forming BIRAC was to support the innovation ecosystem in India, and to nurture innovators from academia and industry to work independently or together," says Shirshendu Mukherjee, mission director of the Program Management Unit at BIRAC. Mukherjee says India has always excelled at basic research but has faced challenges in translating that into commercial outcomes. BIRAC's mission is therefore to "take innovation from the bench to the bedside, from the lab to the field, from the desk to the market", he says.

> "A LOT OF EARLY-STAGE START-UPS ARE GETTING **FUNDED** BUT I THINK THE **CHALLENGE** IS STILL THE LATE STAGE."

In just six years of existence, BIRAC has supported 316 start-ups, which have generated $125 million through 122 products and technologies, including a cattle-feed supplement, a new process to manufacture human albumin and immunoglobulin, microfluidics-based diagnostics and a rapid test for malaria.

Its initiatives include 'biotechnology ignition grants' of up to 5 million rupees for start-ups and entrepreneurs to take a proof-of-concept through to the first major step on the path to commercialization. Another is a 'glue grants' scheme, which connects clinical-science departments with those for basic science in institutes and universities in the hope that this will encourage partnerships and collaborations.

BIRAC has also joined forces with the Bill & Melinda Gates Foundation in Seattle, Washington, on the Grand Challenges India initiative to tackle global health and development problems.

"I always call my Grand Challenges programme 'in India, for India and beyond'," says Mukherjee. "So we will do it in India, we will validate it in India, we will use it India, our citizens will use it, and then if it goes beyond India we are happy to do that."

CONSUMER DEMAND

A similar motivation is driving at least some of the scientists and entrepreneurs such as Acharya, who get into the biotech space because they feel that Western biotechnology isn't necessarily addressing the needs of the Indian population. One example is Vivek Wadhwa, a technology entrepreneur at Harvard Law School in Cambridge, Massachusetts, and at Carnegie Mellon University's College of Engineering at Silicon Valley, California, who has invested in Indian medical-diagnostics company HealthCube in New Delhi.

"I did a big study on the pharmaceutical industry in India, and I concluded that Western companies were not addressing Indian disease because it wasn't profitable enough for them," Wadhwa says.

But as the cost of technologies such as genome sequencing and medical sensors comes down, Wadhwa says, it has now become viable for Indian biotechnologists to harness these advances for the Indian market.

And what a market India is for these innovations. The country's population is 1.36 billion and rising, and health care is one of India's fastest-growing sectors, driven by higher incomes and an increasing prevalence of lifestyle diseases, such as heart disease and stroke. By 2022, the health-care market in India is expected to be worth $372 billion.

"People are finally realizing that the consumer, or the patient, actually has control over their own health," says Acharya. The rising middle class wants better health and medical choices, and she says that's one of the main drivers for investment in biotechnology research and development.

For example, Biocon has developed the first recombinant insulin to be produced in India, and an antibody-based treatment for head and neck cancer. In 2017, Indian vaccine manufacturer Bharat Biotech in Hyderabad began

© 2018 Springer Nature Limited. All rights reserved.



the first clinical trials of its vaccine against the mosquito-borne virus chikungunya, which re-emerged in India in 2006 after 32 years and infected more than 1.4 million people.

Another major driver of the biotechnology boom in India is the accessibility of funding, from both government and private industry. In one 2016 report on biotechnology, India ranked only 49th out of 54 countries. But it scored particularly highly on the availability of venture capital compared to countries such as the United Kingdom, Australia and Canada (see go.nature.com/2rrpuks).

Acharya says that some of the investors who have made their fortunes in manufacturing generic pharmaceuticals are now investing in biotechnology. She says much of the capital investment in early-stage biotechnology is coming from India, whereas investment in medical devices is flowing from Japan, China and the United States. But late-stage investment is still an issue.

"A lot of early-stage start-ups are getting funded but I think the challenge is still the late stage," she says. "It's not just the first two to three years; it's more how do you take it from start-up to scale-up? I think that's the challenge in terms of getting to where we need to get in terms of biotechnology."

HUMAN RESOURCES

One thing India has plenty of is people. Recognizing that human capital can be a key resource for a nation not as well endowed financially as Western countries such as the United States or United Kingdom, the Department of Biotechnology implemented or supported various training initiatives. These include the Biotech Industrial Training Programme, set up in 1993 for recent graduates, and 12 Biotech Finishing Schools in Karnataka state to train Indian graduates and researchers in biotechnology.

That programme "created a very large number of institutions or departments of biotechnology in institutions and also departments of bioinformatics", says Singh. For example, in September, the state of Gujarat proposed India's first university focused entirely on biotechnology.

"A decade or so ago, India didn't have the engineers or scientists it does today — it's been graduating them in droves," says Wadhwa. "It has millions of technologists who now just need to be connected to the medical practice and they can be solving great problems."

Singh notes that these graduates aren't waiting for a job to walk up and tap them on the shoulder; they're taking matters into their own hands. "Graduate students who come out in large numbers from Indian institutes of technology and institutes of management are not looking for jobs so much; they create small start-ups and then they grow very fast," Singh says.

Working in biotechnology in India does present its own unique set of challenges, says Acharya. "Some operational things that you never have to think about in the United States you have to plan more in India, because a lot of times we are still importing the reagents and things like that."

RED TAPE

Although the government of India is enthusiastic about supporting the biotechnology industry, Acharya says the regulatory process for getting products approved could be more streamlined.

In agricultural biotechnology, the government's Genetic Engineering Appraisal Committee has been working to make it easier for companies to get approval for genetically modified crop field trials from state governments.

The drug approvals process in India has hit some rough patches in recent years, and the authors of a 2017 World Health Organization report suggested that innovation there could be outpacing regulation (see go.nature.com/2pkysow). Even the government's own National Biotechnology Development Strategy for 2015–20 acknowledges that timelines and regulatory steps for biotechnology drug approvals are not user-friendly. It has proposed reforms, including the establishment of regulatory departments that are fluent in good practice in the clinical, manufacturing and laboratory arenas.

There are also concerns about the environmental impact of India's pharmaceutical industry. An investigation in 2016 found "unprecedented" levels of pharmaceutical pollution in the water system of Hyderabad (C. Lübbert *et al. Infection* **45**, 479–491; 2017), which is home to a significant proportion of biotech start-ups and generics manufacturers. However, as the US Food and Drug Administration reportedly steps up inspections of overseas pharmaceutical suppliers, environmental standards could be forced to improve.

Despite the challenges, there is palpable excitement about what lies ahead. "Right now, we are seeing the beginnings of a revolution in biotechnology in India," Wadhwa says.

Acharya is still fired with the same enthusiasm that propelled her into biotechnology nearly two decades ago. "Any innovation in this space can actually impact lives," she says. "That's why I continue to be in it." ■

Bianca Nogrady *is a science writer in Sydney, Australia.*

nature jobs

INTERVIEWS
RECRUITMENT
NEWS
OPINION
INTERACT
CAREERS
Q&A
ADVICE
COMMENT
UPDATES
TIPS

blogs.nature.com/naturejobs



A53746

© 2018 Springer Nature Limited. All rights reserved.

nature INDEX 2018 CHINA

NATURE, VOL. 564, ISSUE NO. 7735 (13 DECEMBER 2018)

ALL EYES ON THE PRIZE

China's rise is the story of the century in science. The news this year that China surpassed the United States as the world's largest producer of scientific articles in 2016 should have come as no surprise. Scientific research is central to President Xi Jinping's dream of China becoming an innovation-driven, knowledge-led economy by 2050.

But despite a 75% rise since 2012, China remains a distant second to the US in its overall output in the Nature Index. This suggests it still has a way to go on research quality, because the Nature Index measures the relative contribution of authors to articles published in 82 natural science journals, chosen by a panel of experts as the world's best.

This supplement focusses on selected disciplines that are leading the way to overcoming China's scientific quality deficit. Notably, China is already at, or very near the top of the index rankings in chemistry and plant biology, two areas in which it has arguably the richest research tradition. The story of biomedical engineering shows the country adapting to international standards on the path towards global science leadership, while the cases of astronomy and nanotechnology highlight its willingness to invest the necessary resources.

There are many problems yet to solve, though, if the dream is to become reality. As a recent survey of Chinese academics highlights (S59), and Futao Huang further explains (S71), a fixation on instant success, job insecurity, and the demands of bureaucracy are among hindrances in the research environment to the flourishing of innovative practice.

Catherine Armitage
Chief editor, Nature Index

Ongoing challenge
S58: China publishes more scientific research than the US but still falls short on quality.

Yielding results to feed a people
S62: Rice studies are central to China's strength in plant biology.

China's place among the stars
S64: The FAST telescope dish will thrust radio astronomers into a role of global leadership.

Small science grows large in new hands
S65: China's nanotech industry is forging ahead with funding, talent and international experience.

Engineering a biomedical revolution
S66: Permissive regulations and a pragmatic approach has seen China's bioscience soar.

Strong spending compounds chemistry prowess
S68: The discipline's historic prominence is underpinned by sustained funding and its value to industrial processes.

Quality deficit belies the hype
S70: The pressure for rapid output prevails in China's research workplaces.

Citations strength begins at home
S70: A propensity to cite compatriots could put China in the lead.

The tables
S72: China's top institutions in the Nature Index.

XINHUA NEWS AGENCY/SHUTTERSTOCK



OLI WINWARD

ON THE COVER
China is reaching for the sky in science with the goal of creating an innovative, knowledge-led economy.

EDITORIAL: Catherine Armitage, Smriti Mallapaty, Rebecca Dargie, Gemma Conroy, Herb Brody, Stephen Pincock **ANALYSIS:** Aaron Ballagh, Bo Wu, Willem Sijp **ART & DESIGN:** Tanner Maxwell, Chika Takeda, Madeline Hutchinson, Wojtek Urbanek, Wesley Fernandes **PRODUCTION:** Kay Lewis, Karl Smart, Ian Pope, Nick Bruni, Bob Edenbach, Joern Ishikawa **MARKETING:** Stacy Best Ruel, Angelica Sarne **SALES & PARTNER CONTENT:** Helen Hill, Janet Cen, Ruffi Lu, Stella Yan, Nicole Yu, Sharon Wang, Yingying Zhou, Sara Phillips, Chika Takeda, Aimee Sics, Chris Gilloch **PUBLISHING:** Rebecca Jones, Richard Hughes, David Swinbanks.

NATURE INDEX 2018 CHINA Nature Index 2018 China, a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. This publication is based on the Nature Index, a database that feeds the website, which is maintained and made freely available at natureindex.com.

NATURE EDITORIAL OFFICES
The Campus, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

CUSTOMER SERVICES
To advertise with the Nature Index, please visit natureindex.com/clientservicesfeedback@nature.com
Copyright © 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

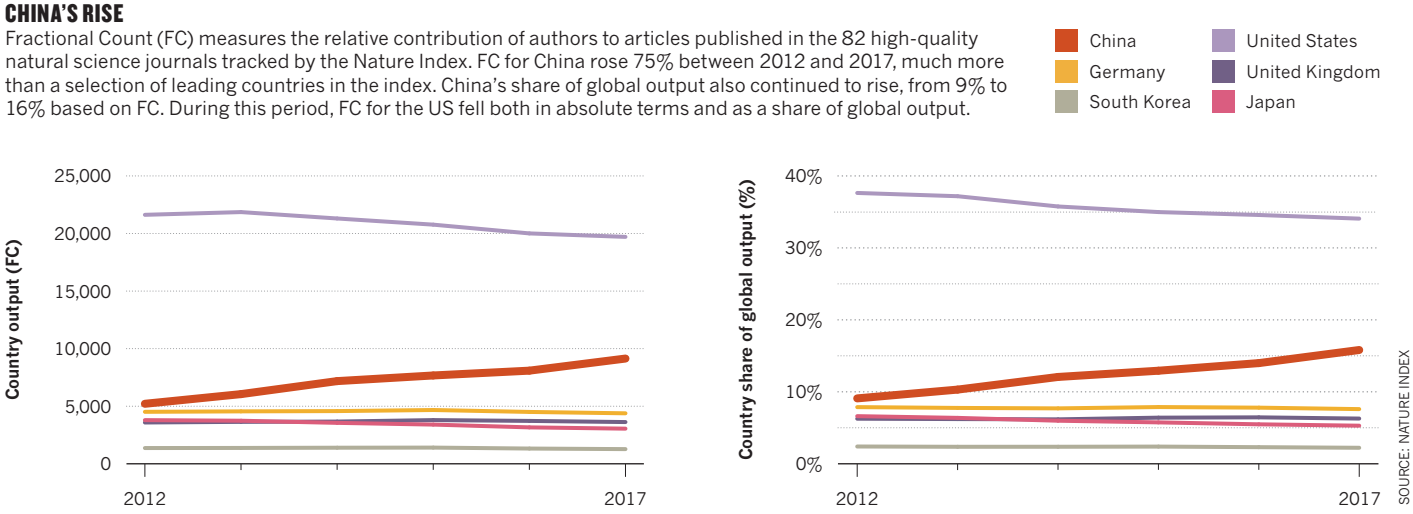
© 2018 Springer Nature Limited. All rights reserved.

ONGOING CHALLENGE

China now publishes more scientific research than the United States, but on measures of quality, including publication of articles in the top-notch journals tracked by the Nature Index, it still falls short, with some notable subject exceptions.

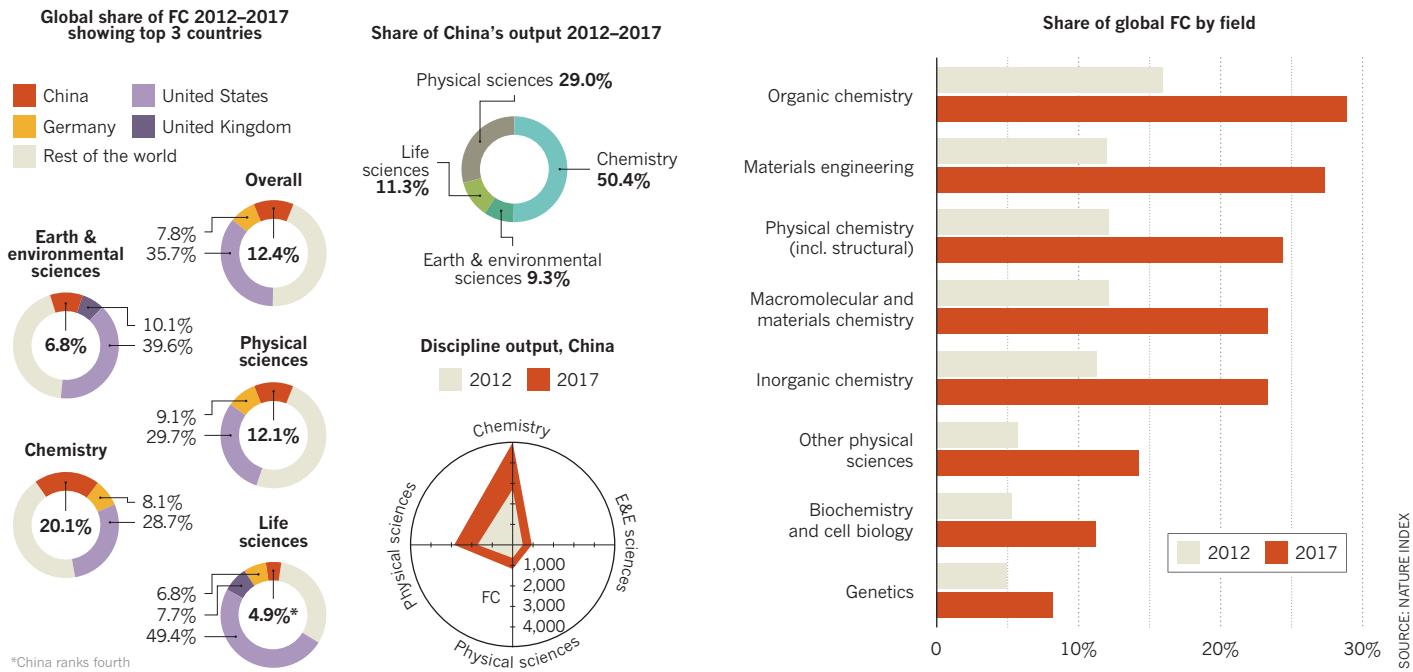
CHINA'S RISE

Fractional Count (FC) measures the relative contribution of authors to articles published in the 82 high-quality natural science journals tracked by the Nature Index. FC for China rose 75% between 2012 and 2017, much more than a selection of leading countries in the index. China's share of global output also continued to rise, from 9% to 16% based on FC. During this period, FC for the US fell both in absolute terms and as a share of global output.



STRENGTH IN NUMBERS

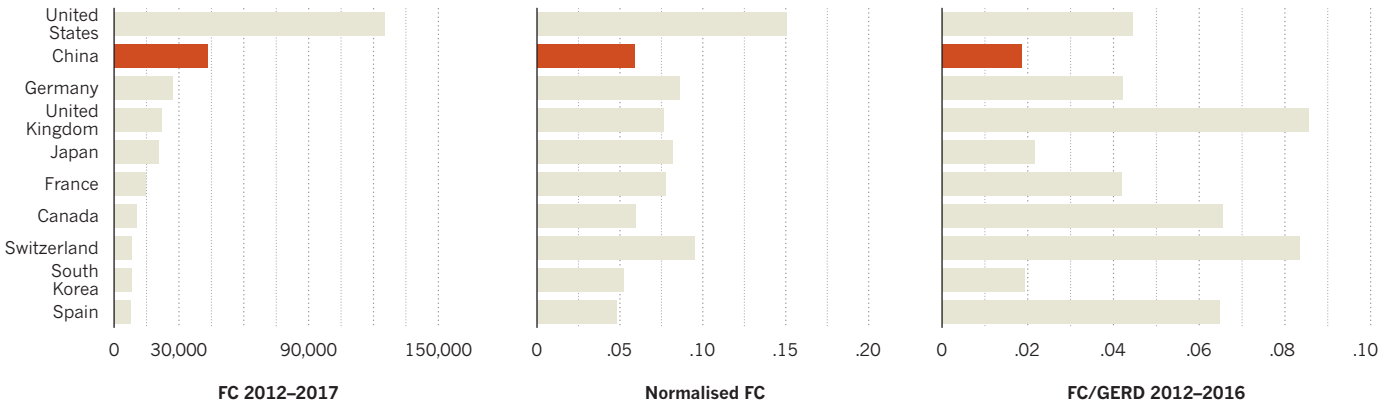
Half of China's FC in the Nature Index concerns chemistry, which is by far the country's strongest field of research in the natural sciences. In the five years to 2017, China produced a fifth of global chemistry output in the Nature Index, but only 4.9% of high-quality output in the life sciences.



© 2018 Springer Nature Limited. All rights reserved.

A LONG WAY TO THE TOP

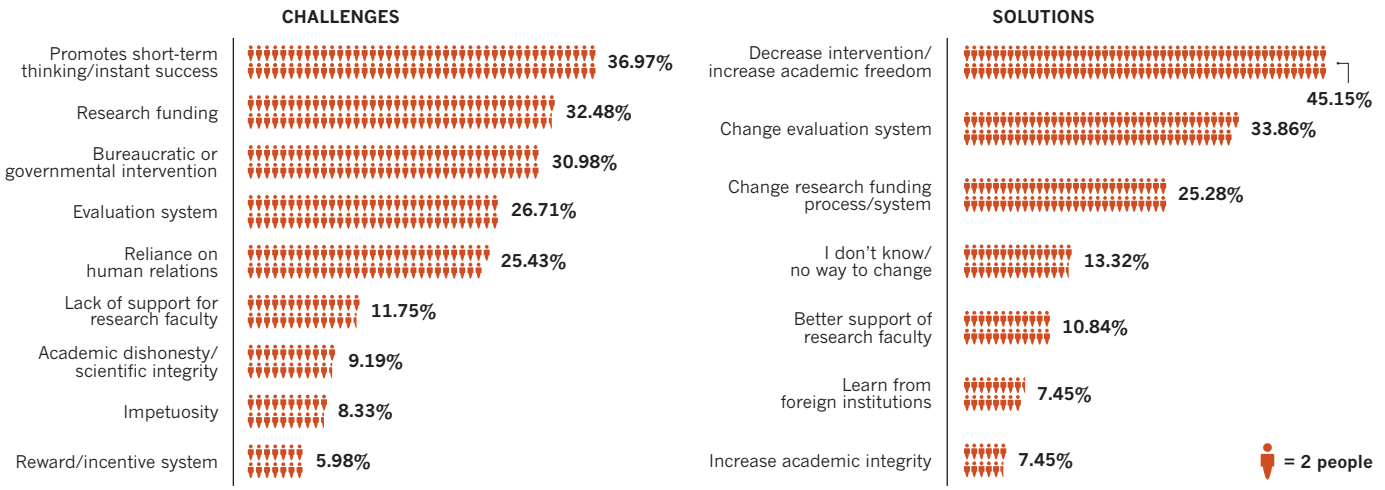
Its total output in the index is second highest, but on measures of quality and efficiency – such as high-quality output normalised against total natural sciences output in the Dimensions database (Normalised FC) and against gross expenditure on R&D, shown here – China is well down the country ranks.



SOURCE: NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE/OECD

VIEW FROM THE BENCH

Short-term thinking and official intervention were high among respondents' concerns in 2016, when two US researchers surveyed 18,000 science, technology, engineering and mathematics researchers in China's top universities. There were 466 responses on challenges and 443 on solutions.



SOURCE: HAN XUEYING AND RICHARD P. APPLEBAUM, *PLOS ONE* 13(4): E0195347 (2018).

PARTNERS IN SCIENCE

Just under 50% of China's articles in the Nature Index were internationally collaborative in 2015–17, about the same proportion as the US, but much less than the UK and Germany (each around 75% internationally collaborative). Here, multilateral collaboration score (MCS) measures the Chinese institution's collaboration with multiple overseas institutions, while the bilateral collaboration score (CS) is a measure of the collaboration between the two institutions shown.

TOP 10 CHINESE INSTITUTIONS BY INTERNATIONAL COLLABORATION 2015-17

CHINESE INSTITUTION	INTERNATIONAL MCS 2015–2017	TOP GLOBAL PARTNER	COUNTRY	PARTNER'S COUNTRY RANK BY FC 2015–2017	CS 2015–2017
Chinese Academy of Sciences	1657.88	Georgia Institute of Technology	United States	35	157.75
Peking University	461.42	Harvard University	United States	1	36.26
Tsinghua University	436.79	Stanford University	United States	2	47.03
Nanjing University	389.25	Stanford University	United States	2	23.41
Zhejiang University	332.59	Nanyang Technological University	Singapore	1	23.05
University of Science and Technology of China	308.91	Nanyang Technological University	Singapore	1	33.73
Fudan University	298.54	Harvard University	United States	1	26.57
Shanghai Jiao Tong University	234.34	Dresden University of Technology	Germany	16	19.95
Xiamen University	200.59	National Institutes of Health	United States	5	17.91
Soochow University	185.72	Florida State University	United States	84	13.05

SOURCE: NATURE INDEX

© 2018 Springer Nature Limited. All rights reserved.

YIELDING RESULTS TO FEED A PEOPLE

Studies to improve the productivity, resistance and taste of rice crops are central to China's commanding position in plant biology.



OLI WINWARD

BY HEPENG JIA

Her students call her Nüshen, which translates as 'goddess', and no wonder. Rice geneticist, Wang Shaokui, was promoted to full professorship on the strength of a single paper, the results of her PhD research published in *Nature Genetics*.

That's close enough to a superhuman feat in China's academic system, where intense competition for tenured positions has created a rampant 'publish or perish' culture among the lower ranks.

Wang's 2012 paper identified a new, highly powerful, rice functional gene, *OsSPL16*, which can control the size, shape and quality of the grain, distinguishing itself from other known genes that control only one of these traits. Her promotion to a full professor in 2014 at Guangzhou-based South China Agricultural University (SCAU), two years after her doctoral graduation there, was a testament not just to her achievement, but the prominence of the plant biology field in China. Scholars in the discipline, including Wang, also hope her trajectory bodes well, by signalling willingness to value quality over quantity in fields where China has sufficient confidence in its home-grown research capacity.

"Now at least in the field of rice biology, there is more emphasis on the role of truly important breakthrough studies of high-quality than on indicators, such as the number of papers, or even the impact factors of the journals that publish these papers," Wang says.

"China has firmly established world leadership in rice science and is steaming ahead, snatching global runner-up positions in biological studies on many other crops," says Yan Jianbing, who specializes in corn research at Wuhan-based Huazhong Agricultural University (HZAU). He cites a Chinese paper examining China's research output in rice, wheat and corn studies, which has been accepted for publication early next year.

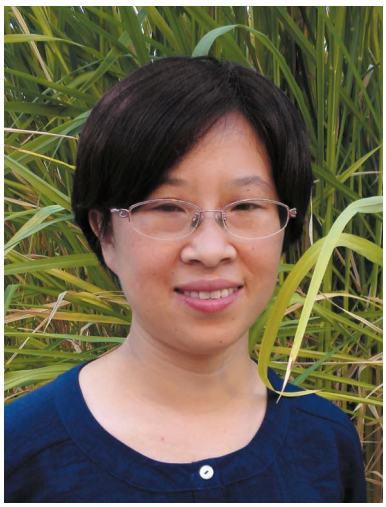
The bibliometric study, of which Yan's HZAU colleague, Liu Bin, is first author, analysed worldwide publications in 31 leading plant science journals indexed by Web of Science. The authors found that between 2012 and 2016, more than half of all the studies on rice in the sampled journals came from China.

A surge in high-quality publications since 2012 by Chinese plant scientists has propelled the country to second place behind the United States in plant biology in the Nature Index, with its fractional count (FC) rising from 47 in 2012 to 90 in 2017.

Based on the growth rates, it is possible that China will overtake the US to become the world's leader in plant biology in the Nature Index within the next six years.

© 2018 Springer Nature Limited. All rights reserved.





Rice geneticist Wang Shaokui, whose genetic studies of rice led to her promotion to full professor.

RUSHING IN

Agricultural research is well-funded in China, reflecting the priority placed on food security. In addition to regular basic research grants from the National Natural Science Foundation of China, there are many megaprojects in agricultural research. For example, in 2008, the Chinese government launched a major initiative on genetically modified crop research coordinated by the Ministry of Agriculture and Rural Affairs, which will receive 24 billion yuan (US$3.5 billion) by 2020. In addition, in 2016, the Ministry of Science and Technology launched a separate key research programme for breeding seven crops, pouring 1.6 billion yuan into 19 research projects for the next five years.

Chinese universities such as Peking, Sun Yat-sen and Zhengzhou are rushing to set up or re-establish their schools of agriculture, which were spun off in the 1950s to form independent agricultural universities across the country.

All newly established agricultural schools have set out to modernize agriculture with breakthroughs in plant science research, and other areas such as low-carbon technologies, and artificial intelligence. In setting them up, many universities also seek to put themselves in the running for extra funding under the next round of World-Class Discipline rankings to be announced under the Double World-Class project in 2022, an initiative to raise the global standing of Chinese universities.

DELICIOUS AND CHEAP

China leads research in rice because the crop is of crucial importance to its food security, says Yan, who is professor and dean of the College of Plant Science and Technology at HZAU. The country has a long tradition in the field, and is home to the world's largest rice research community, with an estimated 3,000 labs and 50,000 scientists nationwide, according to Wang. Access to the latest technologies, particularly new generations of genome sequencers, and well-preserved high-quality samples of diversified rice varieties, point to further significant progress, says Wang.

Studies to examine the molecular mechanisms underlying the productivity and properties of high-yield hybrid rice, as well as its resistance to disease and environmental challenges, have been the focus of Chinese plant biology research in recent years, as set out in the annual reviews of the *Chinese Bulletin of Botany*.

For example, a study led by Han Bin of the Shanghai Institutes for Biological Sciences (SIBS), published in *Nature* in 2016, analysed the genomes of 17 hybrid rice crosses to reveal the genetic mechanism underlying hybrid vigour.

The study, selected by leading scientists for the state-owned publication *Science & Technology Daily* as one of the 10 best breakthroughs nationally, was ranked in importance next to the development by Chinese scientists of hybrid rice in the 1970s. SIBS was China's leading institution in plant biology in the index in 2015–17.

Hybrid rice developed by Chinese scientists in the 1960s and 1970s has fed millions in countries like India, Vietnam, Pakistan and the Philippines and won Yuan Longping, its main developer, the 2004 World Food Prize. Contemporary plant biologists in China, including Wang and Yan, believe their work will also benefit the world by leading the way to new high-yield, high-nutrition and pest-resistant crop varieties.

"Previously, the central policy consideration was to provide enough food, but now, improving the taste of rice and developing diversified food supplies from wheat, corn and soya have become equally important," says Wang.

Wang, by then a professor, received a prestigious 2015 Young Changjiang Scholar award from the Chinese Ministry of Education. In the same year, she published a second study in *Nature Genetics* that identified a gene, *GW7*, that controls rice traits and texture and simultaneously improves yield and grain quality.

Susan McCouch, a rice geneticist at Cornell University in the US, in whose lab Wang had worked, was quoted in *Nature* saying the implications were "enormous": "The rice-breeding community has had this problem, they have been able to improve yield or quality of rice, but almost never simultaneously." McCouch said, according to *Nature*, that in a country where many people eat three rice meals a day, "it will bring pleasure to some of the world's poorest people." ■

PLANT BIOLOGY

TOP INSTITUTIONS 2015–2017:
1. Shanghai Institutes for Biological Sciences (Fractional count: 17.75)
2. Peking University (FC: 17.03)
3. China Agricultural University (FC: 11.64)

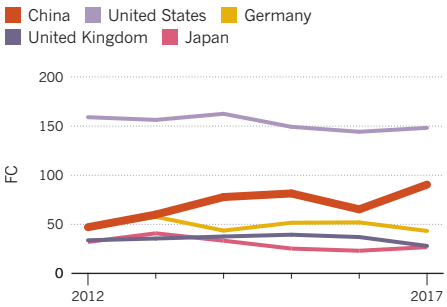
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 6

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 63.1%

TOP CHINESE–INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
Shanghai Institutes for Biological Sciences – Purdue University, United States (CS*: 6.75, Article count: 12)

ATTAINS WORLD NUMBER ONE: 2024**

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score **Extrapolation only

NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

SANDY HARRINGTON

© 2018 Springer Nature Limited. All rights reserved.



Once in use, the scientsists behind FAST intend for it to be as open as possible to use by researchers around the world.

XINHUA NEWS AGENCY/SHUTTERSTOCK

CHINA'S PLACE AMONG THE STARS

The FAST telescope dish, stretching half a kilometre, will thrust China's radio astronomers into a role of global leadership.

BY MARK ZASTROW

Nestled amongst the karst mountains of southwestern China lies a gargantuan spherical dish that fills a valley — the world's single largest radio telescope. Completed in 2016, the Five-hundred-meter Aperture Spherical Radio Telescope (FAST) is being fine-tuned and will soon be fully operational.

The 1.2 billion-yuan (US$180 million) project is a bold investment for China, and puts its radio astronomers in an unfamiliar position — running a world-class facility, with the international community queuing to use it. The role is a welcome one, but not without challenges.

"Relatively speaking, our team is not very experienced," says Li Di, FAST's chief scientist and leader of the radio astronomy division of its operator, the National Astronomical Observatories of China (NOAC) in Beijing.

Radio astronomers are able to see a sky that remains invisible to their optical counterparts — awash in waves from astronomical objects as diverse as spinning neutron stars, black holes, and leftover whispers of the Universe's early inflation. Their discipline is enhancing understanding of the Universe from the Big Bang to galactic structure. The technology used to receive and amplify radio waves from outer space also has practical applications: it was crucial to the development of WiFi, and is driving developments in advanced data processing.

But due to weakness of the signals emitted by the phenomena it studies, radio astronomy is heavily reliant on a few vast facilities with large signal collecting areas. Astronomers also collate results from dishes around the world to improve resolution. This inter-dependence has led the field to embrace international cooperation even faster than its optical counterparts, says Li. It also presents FAST with the opportunity to make an instant global impact.

In recent years, China has surpassed the United Kingdom to become the world's second most prolific nation in high-quality astronomy and space research, trailing only the United States in the index in 2017. Despite limited observing facilities, it has excelled in theoretical solar physics and the complex simulations of the magnetic fields of the Sun and its planets.

But China's radio astronomers have long yearned for larger facilities to make a greater impact, says Peng Bo, FAST's deputy manager and acting observatory director. He recalls going to conferences in the 1980s when China's largest radio dish, 25 metres, was dwarfed by the 100-metre Radio Telescope Effelsberg in Germany. "I felt China's voice was too weak," he says. "Our radio astronomy was far behind the world."

FAST's capabilities will jumpstart the careers of Chinese researchers and students who will have easier access to a world-class facility, says Peng. "In the past, we spent six months abroad to use other countries' telescopes," he says.

Peng and Li were key players in a group of Chinese astronomers who developed the concept of FAST in the mid-1990s. Originally, it was part of a proposal to host the colossal

© 2018 Springer Nature Limited. All rights reserved.

international radio astronomy project now known as the Square Kilometre Array. The SKA collaboration chose sites in South Africa and Australia, and a design using smaller, cheaper dishes, but China's government stepped up in 2007 to build FAST anyway.

Construction on the facility was completed in 2016 and it is still in the commissioning phase. But, it has already found 44 pulsars — neutron stars formed in a supernova explosion with powerful magnetic fields that appear as pulses of radio waves as they spin. FAST could double the known pulsar tally of 2,000, says Peng. It's also ideal for mapping gas clouds between stars and for listening for signals from any alien civilizations.

MIND THE GAP

The next challenge is to wrap up FAST's commissioning and begin operations in earnest — hopefully by 2019. FAST's unique dish is made of 4,450 panels, some of which can be angled by actuators — winches dug into the mountain sides to adjust the telescope's focus. By reshaping its surface, it can be more sensitive to signals from a wider range of direction, off its main axis. But maintaining precise control of them has proven difficult.

"Astronomy is easy; actuators are hard," says Li, who also served as FAST's deputy chief engineer. "And money is harder."

A quirk of China's regulations means FAST is in a funding lull: its construction funds are exhausted, but until it passes its final review from the National Development and Reform Commission, it's not eligible for operational grants and is relying on stopgap funding from the NOAC and its parent organization, the Chinese Academy of Sciences (CAS). This means that the project has not yet received any funds to prepare for the immense amounts of data the telescope will collect. The team has been working to get preliminary results, but the bulk of the data-processing pipeline, which will be crucial to outside users, is yet to be developed.

FAST's leaders remain confident it will pass its project review within a year, at which point the telescope will be deemed a fully fledged national observatory. The design specifications have been met in major areas, such as sensitivity and tracking, says Li. And if it passes, some outside users will be able to access the telescope as soon as late 2019 — if they are willing to shoulder some of the burden of processing the data.

OPEN SKIES

The number of international astronomers that will be among those outside users is yet to be determined. FAST's top scientists intend for it to be as open as possible. "We will strive for an open sky policy because that has been the convention, particularly for radio astronomy, if not all astronomy," says Li. However, the decision lies with NOAC and CAS.

Peng says some have suggested reserving 60–70% of FAST's observing time for Chinese researchers and 20% for foreigners, with 10–20% allocated at the director's discretion.

Li says that the US National Science Foundation has approached Chinese agencies in an attempt to secure access for US scientists. He adds: "We have benefitted from access to US astronomical facilities," as well as data from NASA space telescopes. "The trend is to become more open and that will also help the productivity of the facility."

Peng foresees a challenge in opening the facility to international observers: translating FAST's voluminous documentation for operating and data processing from Chinese to English.

When it does open to other users, there will be no shortage of eager applicants. "It's going to be a great instrument," says Jason Hessels, an astronomer at the Netherlands Institute for Radio Astronomy in Amsterdam. He hopes to point FAST towards recent supernova explosions to search for the faint emission from potential neutron stars. The search could yield a new class of objects — neutron stars emitting bursts of radio waves due not to their rotation, but from the decay of their magnetic fields. "We've never studied such newly born neutron stars before, so there are bound to be surprises." ■

ASTRONOMICAL AND SPACE SCIENCES

TOP INSTITUTIONS 2015–2017:
1. Peking University (Fractional count: 9.12)
2. Fudan University (FC: 8.17)
3. Beijing Normal University (FC:7.80)

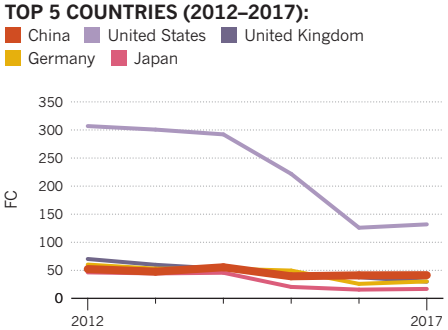
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 1

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 58.5%

TOP CHINESE–INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
Fudan University – University of Tübingen, Germany (CS*: 4.92, Article count: 6)

ATTAINS WORLD NUMBER ONE: NA

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score

SMALL SCIENCE GROWS LARGE IN NEW HANDS

China's nanotech industry is forging ahead, with high funding levels, a maturing talent pool, and international experience.

BY SARAH O'MEARA

Three years ago, chemist David Leigh was ambushed by a smiling master's student from Fudan University waving a CV and looking for a PhD position at Leigh's lab in the United Kingdom.

Leigh, who leads a world-renowned lab at the University of Manchester, had just given a talk on molecular structures in Shanghai. But he was not familiar with Fudan. "I know now it's one of the top universities in China, but I'll admit I didn't then," he says.

The chemist, a fellow of the Royal Society, has become used to the huge enthusiasm exhibited by early-career Chinese scientists who attend his lectures in China. "There'll be crowds at 8.30 a.m. and students gather afterwards for autographs." The CV-brandishing Zhang Liang stood out. Keen to study overseas, he knew winning one of the 10 places available each year at Leigh's prestigious lab was a long shot, so he pushed himself forward.

Happily, it worked out. Leigh, who recently made the world's first molecular robot constructed of atoms, hired Zhang. Two years later, the young Chinese scientist was working on his organic chemistry PhD in Manchester, while managing the creation of Leigh's first nanotech lab in China at East China Normal University (ECNU) in Shanghai.

Over the past 20 years, China has become an influential force in nanoscience research, with particularly high publication rates in catalysis and nanomedicine. Yet despite the productivity and substantial funding, ground-breaking research with broad applications delivering substantial returns on investment has not followed. To deliver such results, researchers say the government needs to invest more in closing the gap between basic research and the industrialization of nanotechnologies. Compared with other big research countries, the percentage of papers with industry co-authors in China remains relatively low.

CATALYSING SCIENCE

China's rise in nanoscience is due to the country's high levels of investment in the field and ambitious targets for research and development overall, such as making the country a

SOURCE: NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

© 2018 Springer Nature Limited. All rights reserved.

international radio astronomy project now known as the Square Kilometre Array. The SKA collaboration chose sites in South Africa and Australia, and a design using smaller, cheaper dishes, but China's government stepped up in 2007 to build FAST anyway.

Construction on the facility was completed in 2016 and it is still in the commissioning phase. But, it has already found 44 pulsars — neutron stars formed in a supernova explosion with powerful magnetic fields that appear as pulses of radio waves as they spin. FAST could double the known pulsar tally of 2,000, says Peng. It's also ideal for mapping gas clouds between stars and for listening for signals from any alien civilizations.

MIND THE GAP

The next challenge is to wrap up FAST's commissioning and begin operations in earnest — hopefully by 2019. FAST's unique dish is made of 4,450 panels, some of which can be angled by actuators — winches dug into the mountain sides to adjust the telescope's focus. By reshaping its surface, it can be more sensitive to signals from a wider range of direction, off its main axis. But maintaining precise control of them has proven difficult.

"Astronomy is easy; actuators are hard," says Li, who also served as FAST's deputy chief engineer. "And money is harder."

A quirk of China's regulations means FAST is in a funding lull: its construction funds are exhausted, but until it passes its final review from the National Development and Reform Commission, it's not eligible for operational grants and is relying on stopgap funding from the NOAC and its parent organization, the Chinese Academy of Sciences (CAS). This means that the project has not yet received any funds to prepare for the immense amounts of data the telescope will collect. The team has been working to get preliminary results, but the bulk of the data-processing pipeline, which will be crucial to outside users, is yet to be developed.

FAST's leaders remain confident it will pass its project review within a year, at which point the telescope will be deemed a fully fledged national observatory. The design specifications have been met in major areas, such as sensitivity and tracking, says Li. And if it passes, some outside users will be able to access the telescope as soon as late 2019 — if they are willing to shoulder some of the burden of processing the data.

OPEN SKIES

The number of international astronomers that will be among those outside users is yet to be determined. FAST's top scientists intend for it to be as open as possible. "We will strive for an open sky policy because that has been the convention, particularly for radio astronomy, if not all astronomy," says Li. However, the decision lies with NOAC and CAS.

Peng says some have suggested reserving 60–70% of FAST's observing time for Chinese researchers and 20% for foreigners, with 10–20% allocated at the director's discretion.

Li says that the US National Science Foundation has approached Chinese agencies in an attempt to secure access for US scientists. He adds: "We have benefitted from access to US astronomical facilities," as well as data from NASA space telescopes. "The trend is to become more open and that will also help the productivity of the facility."

Peng foresees a challenge in opening the facility to international observers: translating FAST's voluminous documentation for operating and data processing from Chinese to English.

When it does open to other users, there will be no shortage of eager applicants. "It's going to be a great instrument," says Jason Hessels, an astronomer at the Netherlands Institute for Radio Astronomy in Amsterdam. He hopes to point FAST towards recent supernova explosions to search for the faint emission from potential neutron stars. The search could yield a new class of objects — neutron stars emitting bursts of radio waves due not to their rotation, but from the decay of their magnetic fields. "We've never studied such newly born neutron stars before, so there are bound to be surprises." ■

ASTRONOMICAL AND SPACE SCIENCES

TOP INSTITUTIONS 2015–2017:
1. Peking University (Fractional count: 9.12)
2. Fudan University (FC: 8.17)
3. Beijing Normal University (FC:7.80)

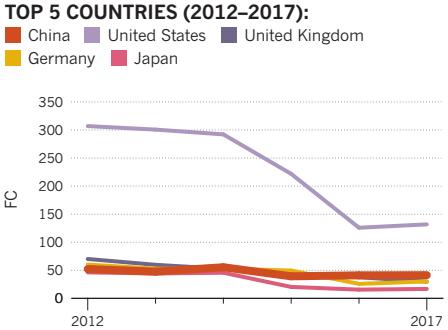
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 1

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 58.5%

TOP CHINESE–INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
Fudan University – University of Tübingen, Germany (CS*: 4.92, Article count: 6)

ATTAINS WORLD NUMBER ONE: NA

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score

SMALL SCIENCE GROWS LARGE IN NEW HANDS

China's nanotech industry is forging ahead, with high funding levels, a maturing talent pool, and international experience.

BY SARAH O'MEARA

Three years ago, chemist David Leigh was ambushed by a smiling master's student from Fudan University waving a CV and looking for a PhD position at Leigh's lab in the United Kingdom.

Leigh, who leads a world-renowned lab at the University of Manchester, had just given a talk on molecular structures in Shanghai. But he was not familiar with Fudan. "I know now it's one of the top universities in China, but I'll admit I didn't then," he says.

The chemist, a fellow of the Royal Society, has become used to the huge enthusiasm exhibited by early-career Chinese scientists who attend his lectures in China. "There'll be crowds at 8.30 a.m. and students gather afterwards for autographs." The CV-brandishing Zhang Liang stood out. Keen to study overseas, he knew winning one of the 10 places available each year at Leigh's prestigious lab was a long shot, so he pushed himself forward.

Happily, it worked out. Leigh, who recently made the world's first molecular robot constructed of atoms, hired Zhang. Two years later, the young Chinese scientist was working on his organic chemistry PhD in Manchester, while managing the creation of Leigh's first nanotech lab in China at East China Normal University (ECNU) in Shanghai.

Over the past 20 years, China has become an influential force in nanoscience research, with particularly high publication rates in catalysis and nanomedicine. Yet despite the productivity and substantial funding, ground-breaking research with broad applications delivering substantial returns on investment has not followed. To deliver such results, researchers say the government needs to invest more in closing the gap between basic research and the industrialization of nanotechnologies. Compared with other big research countries, the percentage of papers with industry co-authors in China remains relatively low.

CATALYSING SCIENCE

China's rise in nanoscience is due to the country's high levels of investment in the field and ambitious targets for research and development overall, such as making the country a

SOURCE: NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

© 2018 Springer Nature Limited. All rights reserved.

SOURCE: NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

world-leading scientific powerhouse by 2050.

In 1990, the government began investing in nanoscience via schemes such as the State Science and Technology Commission's Climbing Up project on nanomaterials. In 1999, the Ministry of Science and Technology began a basic research project, Nanomaterial and Nanostructure, and by 2006 the field had become one of the four pillars of basic research that received targeted funding from central government. Last year, the Suzhou Institute of Nano-Tech and Nano-Bionics announced a US$200-million plan to build the world's largest multifunctional nanoscience research facility for computer and robot technologies.

China's nano-related output has grown from 820 papers in 1997 to more than 52,000 papers in 2016 in the Science Citation Index. Four of the top 10 institutions for high-quality nanotechnology output are in China, according to the index.

The most popular area of the country's nanoscience papers is in catalysis research, according to the number of articles listed in *Nature*'s Nano database, a web platform that examines the quantity and impact of nano-related research papers published globally. Experts predict this area of nanoscience research will continue to flourish. A team of scientists led by Bao Xinhe at the Dalian Institute of Chemical Physics has developed a catalyst that enables the direct conversion of synthetic fuel gas to light olefins, the basic building blocks of plastics.

NANOTECHNOLOGY

TOP INSTITUTIONS 2015–2017:
1. Tsinghua University (Fractional count: 29.16)
2. Peking University (FC: 28.45)
3. Soochow University (FC: 24.23)

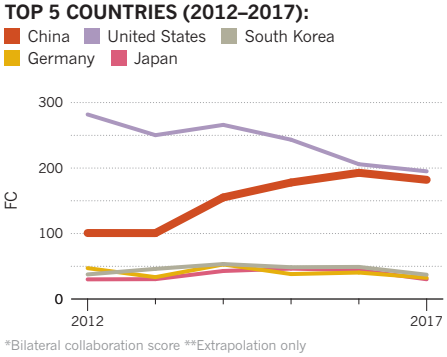
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 7

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 49.9%

TOP CHINESE–INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
Soochow University – Monash University, Australia (CS*: 3.92, Article count: 6)

ATTAINS WORLD NUMBER ONE: 2018**

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score **Extrapolation only

MODEST RETURNS

Chinese researchers have also contributed to nanomedicine applications, such as improved methods for treating cancer. Despite such developments, Dai Qing, who returned to China in 2012 to launch a nanophotonics laboratory at the National Center for Nanoscience and Technology, is among those who believe Chinese scientists should push for stronger returns on investment. "We need to find a stand-out application to demonstrate that it is beneficial to the country to spend this money, not just talk about the possibilities," he says.

Dai says that the grants structure has changed to reflect a push for tangible outcomes. "If you want a flagship grant you definitely need to find an industry partner." Examples of change include developing interdisciplinary teams within Chinese institutes, taking a stronger lead in international projects, and working more closely with industry partners. For example, in May 2018, a minimally invasive cancer therapy was trialled in Shanghai. The 'nano gun' is a device loaded with anti-cancer agents that is injected into tumours. It was developed by a Chinese team working with Algerian researchers, based in France. If trials are successful, it will be developed for application in China.

As a young scientist Zhang, whose drive impressed Leigh, broke with tradition by initiating a conversation with his former professor, Hai-Bo Yang, from ECNU's Department of Chemistry about starting a lab, which is now open. "Historically, respected Chinese academics court international talent to create collaborations. I realized young starters could do it with the help of senior local professors," Zhang says. "To succeed in academia you need to learn to make connections."

In 2018, China established its first private research institute, Westlake University in Hangzhou. It is backed by some of the country's wealthiest industrialists, including Ma Huateng (Pony), founder and CEO of Internet giant Tencent, and Wang Jianlin, founder and chairman of the Dalian Wanda Group.

Many agree that while China has a bright future, cultural factors can hinder its ability to compete. On Zhang's new team are four postdocs, two from China, and two from the UK and Germany. All were invited to Manchester for training. During that period Zhang noticed how culture and language influence research styles. "Discoveries often come from conflict or argument. But Chinese people can be culturally averse to this," he says. "Also, when you're working in your second language, it can be hard to argue your point."

Zhang believes the country will become a nanotech world leader. Certainly, Leigh believes that his days of being revered in China are numbered. ■

ENGINEERING A BIOMEDICAL REVOLUTION

A permissive regulatory climate and a pragmatic approach has seen China's bioscience sector soar.

BY SMRITI MALLAPATY

For the past 20 years, French neuroscientist Erwan Bezard has spent at least one week every two months in Beijing. Bezard makes the long journey from France to visit the primates bred in Chinese labs.

China has become the top destination for research involving these animals, which are invaluable models for studying human disease. Other countries do not breed the primates in such large numbers or to the standard produced in China.

"Some 95% of papers using transgenic monkeys come from China," says Bezard, director of the Institute of Neurodegenerative Diseases at the University of Bordeaux, and manager of his own lab at the Institute of Laboratory Animal Sciences, Chinese Academy of Medical Sciences. Among recent breakthroughs, researchers at the Chinese Academy of Sciences (CAS) have genetically modified cynomolgus monkeys so they exhibit autistic-like behaviours, to better understand what causes the disorder, and how to treat it. CAS scientists have also cloned primates using a technique similar to the one that produced Dolly the sheep. Bezard has used rhesus monkeys to show how brain–computer interfaces can restore leg movement after spinal cord injury.

These developments have coincided with improvements in the regulation and enforcement of international standards in the biosciences in China. Two events were critical to the process: the 2003 SARS outbreak, which put a spotlight on the issue of wildlife and lab animal management, and the creation in China of the world's first human–rabbit embryos in 2001, which provoked an international public relations crisis for the country.

The Chinese government recognizes that bioscience will play a major role in its global competitiveness. Biomedicine, synthetic biology and regenerative medical techniques are listed as strategic fields and industries in China's 13th Five-Year Plan. "China doesn't want to miss the life-science biotech revolution," says Cao Cong, an innovation studies researcher at the University of Nottingham Ningbo China.

Scientists have also realized that to gain

© 2018 Springer Nature Limited. All rights reserved.

SOURCE: NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

world-leading scientific powerhouse by 2050.

In 1990, the government began investing in nanoscience via schemes such as the State Science and Technology Commission's Climbing Up project on nanomaterials. In 1999, the Ministry of Science and Technology began a basic research project, Nanomaterial and Nanostructure, and by 2006 the field had become one of the four pillars of basic research that received targeted funding from central government. Last year, the Suzhou Institute of Nano-Tech and Nano-Bionics announced a US$200-million plan to build the world's largest multifunctional nanoscience research facility for computer and robot technologies.

China's nano-related output has grown from 820 papers in 1997 to more than 52,000 papers in 2016 in the Science Citation Index. Four of the top 10 institutions for high-quality nanotechnology output are in China, according to the index.

The most popular area of the country's nanoscience papers is in catalysis research, according to the number of articles listed in *Nature*'s Nano database, a web platform that examines the quantity and impact of nano-related research papers published globally. Experts predict this area of nanoscience research will continue to flourish. A team of scientists led by Bao Xinhe at the Dalian Institute of Chemical Physics has developed a catalyst that enables the direct conversion of synthetic fuel gas to light olefins, the basic building blocks of plastics.

NANOTECHNOLOGY

TOP INSTITUTIONS 2015–2017:
1. Tsinghua University (Fractional count: 29.16)
2. Peking University (FC: 28.45)
3. Soochow University (FC: 24.23)

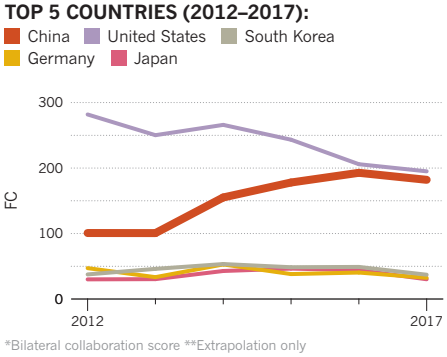
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 7

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 49.9%

TOP CHINESE–INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
Soochow University – Monash University, Australia (CS*: 3.92, Article count: 6)

ATTAINS WORLD NUMBER ONE: 2018**

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score **Extrapolation only

MODEST RETURNS

Chinese researchers have also contributed to nanomedicine applications, such as improved methods for treating cancer. Despite such developments, Dai Qing, who returned to China in 2012 to launch a nanophotonics laboratory at the National Center for Nanoscience and Technology, is among those who believe Chinese scientists should push for stronger returns on investment. "We need to find a stand-out application to demonstrate that it is beneficial to the country to spend this money, not just talk about the possibilities," he says.

Dai says that the grants structure has changed to reflect a push for tangible outcomes. "If you want a flagship grant you definitely need to find an industry partner." Examples of change include developing interdisciplinary teams within Chinese institutes, taking a stronger lead in international projects, and working more closely with industry partners. For example, in May 2018, a minimally invasive cancer therapy was trialled in Shanghai. The 'nano gun' is a device loaded with anti-cancer agents that is injected into tumours. It was developed by a Chinese team working with Algerian researchers, based in France. If trials are successful, it will be developed for application in China.

As a young scientist Zhang, whose drive impressed Leigh, broke with tradition by initiating a conversation with his former professor, Hai-Bo Yang, from ECNU's Department of Chemistry about starting a lab, which is now open. "Historically, respected Chinese academics court international talent to create collaborations. I realized young starters could do it with the help of senior local professors," Zhang says. "To succeed in academia you need to learn to make connections."

In 2018, China established its first private research institute, Westlake University in Hangzhou. It is backed by some of the country's wealthiest industrialists, including Ma Huateng (Pony), founder and CEO of Internet giant Tencent, and Wang Jianlin, founder and chairman of the Dalian Wanda Group.

Many agree that while China has a bright future, cultural factors can hinder its ability to compete. On Zhang's new team are four postdocs, two from China, and two from the UK and Germany. All were invited to Manchester for training. During that period Zhang noticed how culture and language influence research styles. "Discoveries often come from conflict or argument. But Chinese people can be culturally averse to this," he says. "Also, when you're working in your second language, it can be hard to argue your point."

Zhang believes the country will become a nanotech world leader. Certainly, Leigh believes that his days of being revered in China are numbered. ■

ENGINEERING A BIOMEDICAL REVOLUTION

A permissive regulatory climate and a pragmatic approach has seen China's bioscience sector soar.

BY SMRITI MALLAPATY

For the past 20 years, French neuroscientist Erwan Bezard has spent at least one week every two months in Beijing. Bezard makes the long journey from France to visit the primates bred in Chinese labs.

China has become the top destination for research involving these animals, which are invaluable models for studying human disease. Other countries do not breed the primates in such large numbers or to the standard produced in China.

"Some 95% of papers using transgenic monkeys come from China," says Bezard, director of the Institute of Neurodegenerative Diseases at the University of Bordeaux, and manager of his own lab at the Institute of Laboratory Animal Sciences, Chinese Academy of Medical Sciences. Among recent breakthroughs, researchers at the Chinese Academy of Sciences (CAS) have genetically modified cynomolgus monkeys so they exhibit autistic-like behaviours, to better understand what causes the disorder, and how to treat it. CAS scientists have also cloned primates using a technique similar to the one that produced Dolly the sheep. Bezard has used rhesus monkeys to show how brain–computer interfaces can restore leg movement after spinal cord injury.

These developments have coincided with improvements in the regulation and enforcement of international standards in the biosciences in China. Two events were critical to the process: the 2003 SARS outbreak, which put a spotlight on the issue of wildlife and lab animal management, and the creation in China of the world's first human–rabbit embryos in 2001, which provoked an international public relations crisis for the country.

The Chinese government recognizes that bioscience will play a major role in its global competitiveness. Biomedicine, synthetic biology and regenerative medical techniques are listed as strategic fields and industries in China's 13th Five-Year Plan. "China doesn't want to miss the life-science biotech revolution," says Cao Cong, an innovation studies researcher at the University of Nottingham Ningbo China.

Scientists have also realized that to gain

© 2018 Springer Nature Limited. All rights reserved.



JIN LIWANG/XINHUA/EYEVINE

Genetically identical cloned monkeys Zhong Zhong and Hua Hua are the first primates to emerge from the method that produced Dolly the sheep.

global recognition for their achievements, they must play by internationally accepted rules, says Du Yanan, a biomedical engineer at Tsinghua University, who returned to China in 2010 after three years at Harvard-MIT Health Sciences and Technology.

Chinese life scientists have used advanced medical imaging technology to better detect malignant nodules in the lungs, created the first monkeys using the CRISPR–Cas9 gene-editing technique, and discovered that fetal DNA flows through the mother's bloodstream. That advance led to the development of a non-invasive test for Down's syndrome during pregnancy, which is used around the world.

In the Nature Index, China is the second leading contributor to biomedical engineering articles after the United States, measured by its contribution to the authorship of papers in 82 high-quality research journals in 2015–17.

CROSS-EXAMINED

In 2001, such advances were unthinkable. In September of that year, Chinese newspapers reported that Chen Xigu, a scientist at Sun Yat-sen University in Guangzhou, had successfully grown rabbit embryos injected with skin-cell nuclei taken from a seven-year-old boy.

"At that time, no one thought that China could make such a breakthrough," says Joy Zhang, a sociologist at the University of Kent in the United Kingdom. The resulting hybrids could be used to derive human embryonic stem cells, useful for regenerative medicine.

The national response was buoyant, but brief. Within days, there was international outcry over the research and China was being called the 'Wild East' of biology. Startled by the furore, the government effectively banned all hybrid embryonic stem cell research, says Zhang. Chen's hybrid research came to an abrupt end, though he remained on the Sun Yat-sen faculty and continued to supervise students, particularly on somatic (non-reproductive) stem cells.

China's bioethics landscape has evolved since the incident. In 2003, the Ministry of Science and Technology and the then Ministry of Health (MoH) issued guidelines for human embryonic stem cell research. Between 2009 and 2013, the MoH introduced administrative measures and regulations governing the clinical application of medical technology and stem cells. A national standard for lab animal institutions came into effect in 2014.

Zhang describes the government's regulatory approach as pragmatic, in which it "copy-pastes" international regulations, largely following the relatively permissive stance of the UK. By clearing the regulatory route for research, says Zhang, China has attracted many overseas-Chinese scientists and non-Chinese collaborators. "Permissive regulation has helped China's quick ascent," says Zhang.

INFORMED CONSENT

On his return to Beijing in 2010, Du experienced lab-culture shock. Researchers were raising lab animals under varying conditions and killing them without humane procedures. Doctors were handing over patient samples to researchers without patient consent.

These practices, although expedient, come with serious risks. Unscrupulous behaviour opens science to criticism, and can make research ineligible for publication in top journals. And, the failure to follow procedures undermines the reproducibility of the results.

Submissions to many reputable journals must be accompanied by approvals from ethics committees. In the years since his return, Du — who has co-authored several recent papers on methods for introducing stem cells into the body — has witnessed great progress in biomedical research ethics.

For example, the first accreditation of a Chinese facility by the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC International) was in 2006; by 2016 around 60 Chinese programmes were accredited by this organization, a non-profit promoting the responsible care and use of animals in science under a voluntary certification framework. China has also begun to take more initiative in global policy debates

© 2018 Springer Nature Limited. All rights reserved.

and set standards in emerging fields, such as stem cells and synthetic biology.

When in 2015, Chinese researchers became the first to use CRISPR on nonviable human embryos, sparking another global ethics debate, the government's response was much more measured than it had been in the recent past, says Zhang. The government clarified its position and regulatory procedures, specifying that embryo gene editing is permitted in China for basic and preclinical research, but prohibited for clinical or reproductive use.

VAST BACKYARD

But, in ethical practice, Chinese science remains highly variable. Decisions about how to implement broad-brush guidelines are left to the discretion of institutions and researchers. The recent claims of genome-edited twins by He Jiankui at the Southern University of Science and Technology of China (SUSTech), which SUSTech has distanced itself from and more than 100 Chinese biomedical researchers have strongly condemned, is a case in point.

Implementation remains patchy, says Du, especially in universities and hospitals in remote regions. When it comes to enforcing standard practices across the whole of China, he says, "we still have a long way to go."

Part of the problem, says Zhang, is one of communication. Good enforcement, she says, requires not just top-down monitoring, but also engagement with the public about "what they should expect and what they are, by law, entitled to." A failure to interact with public interest groups about the rules facilitates the spread of rumours, misconceptions and distrust in science, says Zhang. ■

BIOMEDICAL ENGINEERING

TOP INSTITUTIONS 2015–2017:
1. Soochow University (Fractional count: 11.39)
2. Nanjing University (FC: 10.07)
3. Xi'an Jiaotong University (FC: 7.71)

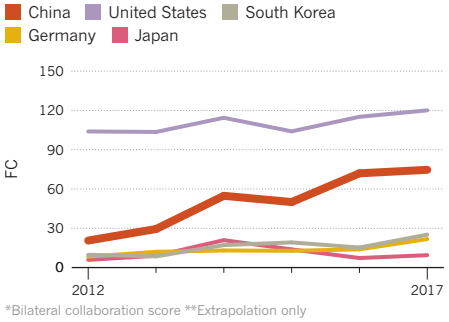
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 5

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 51.7%

TOP CHINESE-INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
Beijing Institute of Nanoenergy and Nanosystems, CAS – Georgia Institute of Technology, United States (CS: 6.27, Article count: 12)

ATTAINS WORLD NUMBER ONE: 2022*

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score **Extrapolation only

STRONG SPENDING COMPOUNDS CHEMISTRY PROWESS

The discipline's historic prominence in China is underpinned by its crucial value to industrial processes. Committed funding sees it leading the way in emerging areas, such as nanomaterials.

BY HEPENG JIA

There were no research labs and no chemical testing devices when Lu Wei and colleagues joined the chemistry department of the new Southern University of Science and Technology (SUSTech) in Shenzhen in 2012. Their only teaching demonstration labs were in a makeshift building; equipment had to be set up and dismantled for each class. "Our graduate students had to cycle for several kilometres to test our samples in partner labs nearby," says Lu, founding head of the department.

As their labs took shape around them, lively discussions stretching into the evenings, sometimes with beer, sparked ideas between young faculty members, recalls Lu. Teams rearranged themselves in new combinations to stretch limited resources further by simplifying logistics.

When state-of-the-art labs were built two years later, with generous funding from the Shenzhen municipal government, the young chemists at SUSTech quickly made breakthroughs, rising to be among the top 50 in China for high-quality chemistry research with a fractional count (FC) of 83.41 for 2015–17 in the Nature Index.

One recent study in *Science*, lead by SUSTech chemist, Tan Bin, identified a catalyst that improves the efficiency of the chemical reaction known as Ugi, which has been widely used to synthesize compounds, especially in the search for new pharmaceuticals.

From 2012 to 2017, China's FC in the Nature Index for chemistry grew by 84% from 2,712 to 4,993, ranking it as the world's second after the United States. By contrast, the US saw a decline of 10% from 6,026 to 5,451.

Among the chemistry sub-disciplines, China surpassed the US for top position in organic chemistry in 2015. It has run second to the US in most other chemistry sub-disciplines in recent years.

"The solid scientific foundation built by Chinese chemists over many years, the nurturing of young talent, and the state's surging investment in research have combined to contribute to this significant growth," says Chen Xiaoming, a chemist at Guangzhou-based Sun Yat-sen University (SYSU), who is also a member of the Chinese Academy of Sciences (CAS).

Chemistry has long been China's top-performing discipline in terms of international research publications. Its crucial importance to many industrial processes has guaranteed it sustained attention since modern science was introduced to the country, says Lu.

ORGANIC GROWTH

Organic chemistry, which studies compounds and materials containing carbon, is the strongest sub-discipline, due to its "wide application, easier operation and huge number of researchers," says Lu, adding that it is very cheap and quick to set up an organic chemistry lab.

There has been rapid growth in studies on organic synthesis methods, nano-catalysts, synthetic organofluorine chemistry, visible-light-driven organic reactions, and natural product chemistry. So said a May 2017 special edition dedicated to organic chemistry of China's most prestigious multidisciplinary journal, *National Science Review*.

Ma Shengming, guest editor of the edition and a leading chemist at CAS Shanghai Institute of Organic Chemistry (SIOC), pointed out the great industrial and environmental benefits of many of these studies which involved interdisciplinary collaboration with emerging areas such as nanomaterials.

For example, according to Hu Jinbo of SIOC, which topped the Nature Index in organic chemistry with an FC of 139.17 for 2015–17, Chinese chemists have made significant progress in improving synthetic organofluorine chemistry, which can reduce chemical pollution. In another example, visible-light-driven organic reactions greatly lower the energy needed for industrial organic chemical synthesis, wrote Wu Lizhu of the CAS Technical Institute of Physics and Chemistry in the special issue.

The large recruitment of chemists particularly through the Young Thousand Talents Plan has provided a fresh boost to the booming study of chemistry in China, says Zhao Dongbing, a chemist at Tianjin-based Nankai University, who was enrolled as a scholar under the plan after completing his postdoctoral research at Cornell University and the University of Washington in Seattle. China launched the plan in 2008 to attract established scientists and high-tech entrepreneurs to return to China. In 2011, the programme was extended to young scholars, mostly

NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

© 2018 Springer Nature Limited. All rights reserved.

and set standards in emerging fields, such as stem cells and synthetic biology.

When in 2015, Chinese researchers became the first to use CRISPR on nonviable human embryos, sparking another global ethics debate, the government's response was much more measured than it had been in the recent past, says Zhang. The government clarified its position and regulatory procedures, specifying that embryo gene editing is permitted in China for basic and preclinical research, but prohibited for clinical or reproductive use.

VAST BACKYARD

But, in ethical practice, Chinese science remains highly variable. Decisions about how to implement broad-brush guidelines are left to the discretion of institutions and researchers. The recent claims of genome-edited twins by He Jiankui at the Southern University of Science and Technology of China (SUSTech), which SUSTech has distanced itself from and more than 100 Chinese biomedical researchers have strongly condemned, is a case in point.

Implementation remains patchy, says Du, especially in universities and hospitals in remote regions. When it comes to enforcing standard practices across the whole of China, he says, "we still have a long way to go."

Part of the problem, says Zhang, is one of communication. Good enforcement, she says, requires not just top-down monitoring, but also engagement with the public about "what they should expect and what they are, by law, entitled to." A failure to interact with public interest groups about the rules facilitates the spread of rumours, misconceptions and distrust in science, says Zhang. ■

BIOMEDICAL ENGINEERING

TOP INSTITUTIONS 2015–2017:
1. Soochow University (Fractional count: 11.39)
2. Nanjing University (FC: 10.07)
3. Xi'an Jiaotong University (FC: 7.71)

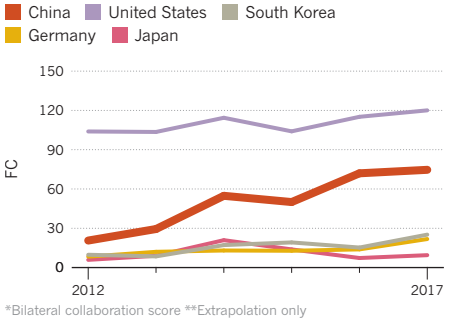
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 5

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 51.7%

TOP CHINESE-INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
Beijing Institute of Nanoenergy and Nanosystems, CAS – Georgia Institute of Technology, United States (CS: 6.27, Article count: 12)

ATTAINS WORLD NUMBER ONE: 2022*

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score **Extrapolation only

STRONG SPENDING COMPOUNDS CHEMISTRY PROWESS

The discipline's historic prominence in China is underpinned by its crucial value to industrial processes. Committed funding sees it leading the way in emerging areas, such as nanomaterials.

BY HEPENG JIA

There were no research labs and no chemical testing devices when Lu Wei and colleagues joined the chemistry department of the new Southern University of Science and Technology (SUSTech) in Shenzhen in 2012. Their only teaching demonstration labs were in a makeshift building; equipment had to be set up and dismantled for each class. "Our graduate students had to cycle for several kilometres to test our samples in partner labs nearby," says Lu, founding head of the department.

As their labs took shape around them, lively discussions stretching into the evenings, sometimes with beer, sparked ideas between young faculty members, recalls Lu. Teams rearranged themselves in new combinations to stretch limited resources further by simplifying logistics.

When state-of-the-art labs were built two years later, with generous funding from the Shenzhen municipal government, the young chemists at SUSTech quickly made breakthroughs, rising to be among the top 50 in China for high-quality chemistry research with a fractional count (FC) of 83.41 for 2015–17 in the Nature Index.

One recent study in *Science*, lead by SUSTech chemist, Tan Bin, identified a catalyst that improves the efficiency of the chemical reaction known as Ugi, which has been widely used to synthesize compounds, especially in the search for new pharmaceuticals.

From 2012 to 2017, China's FC in the Nature Index for chemistry grew by 84% from 2,712 to 4,993, ranking it as the world's second after the United States. By contrast, the US saw a decline of 10% from 6,026 to 5,451.

Among the chemistry sub-disciplines, China surpassed the US for top position in organic chemistry in 2015. It has run second to the US in most other chemistry sub-disciplines in recent years.

"The solid scientific foundation built by Chinese chemists over many years, the nurturing of young talent, and the state's surging investment in research have combined to contribute to this significant growth," says Chen Xiaoming, a chemist at Guangzhou-based Sun Yat-sen University (SYSU), who is also a member of the Chinese Academy of Sciences (CAS).

Chemistry has long been China's top-performing discipline in terms of international research publications. Its crucial importance to many industrial processes has guaranteed it sustained attention since modern science was introduced to the country, says Lu.

ORGANIC GROWTH

Organic chemistry, which studies compounds and materials containing carbon, is the strongest sub-discipline, due to its "wide application, easier operation and huge number of researchers," says Lu, adding that it is very cheap and quick to set up an organic chemistry lab.

There has been rapid growth in studies on organic synthesis methods, nano-catalysts, synthetic organofluorine chemistry, visible-light-driven organic reactions, and natural product chemistry. So said a May 2017 special edition dedicated to organic chemistry of China's most prestigious multidisciplinary journal, *National Science Review*.

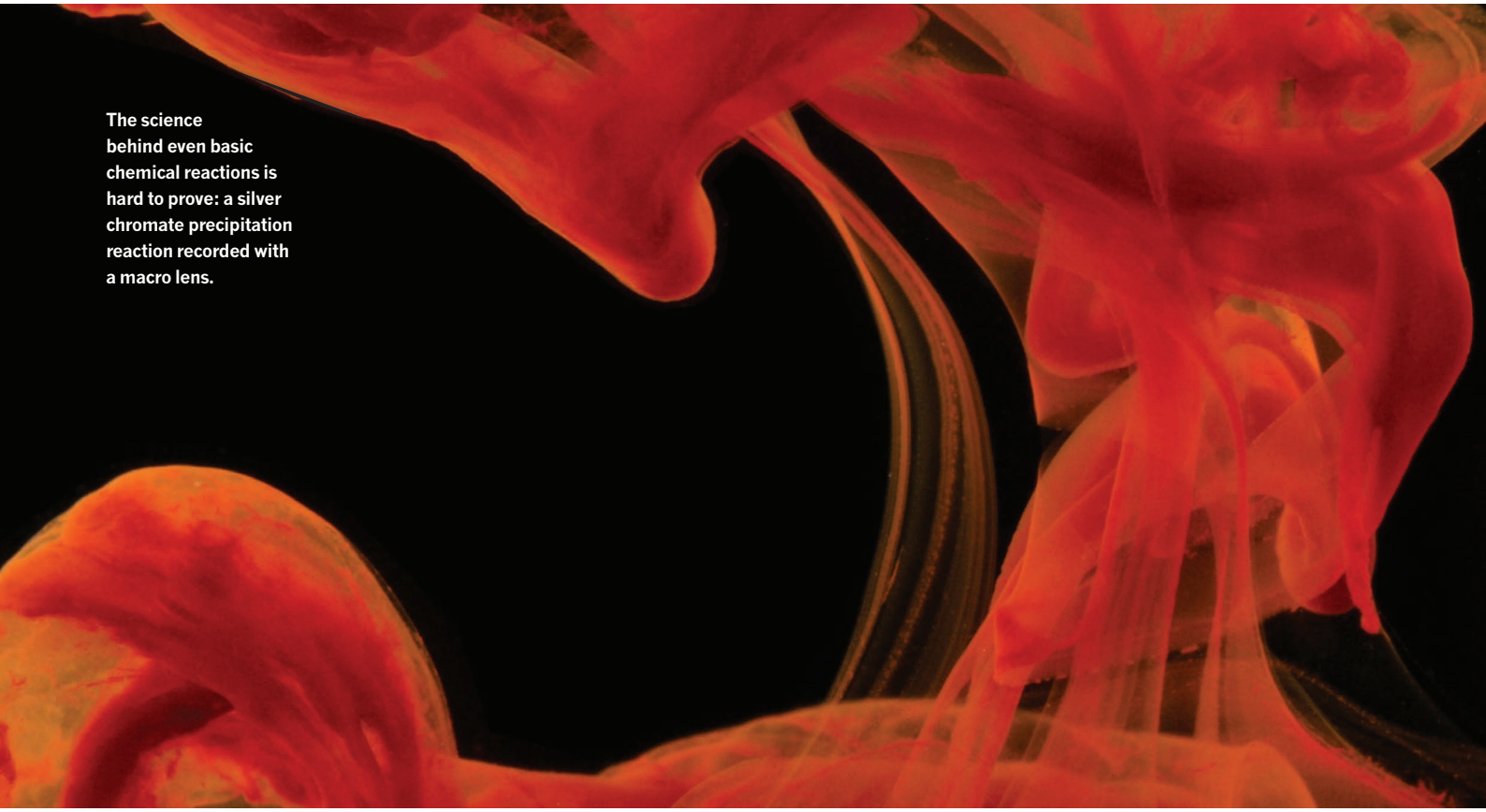
Ma Shengming, guest editor of the edition and a leading chemist at CAS Shanghai Institute of Organic Chemistry (SIOC), pointed out the great industrial and environmental benefits of many of these studies which involved interdisciplinary collaboration with emerging areas such as nanomaterials.

For example, according to Hu Jinbo of SIOC, which topped the Nature Index in organic chemistry with an FC of 139.17 for 2015–17, Chinese chemists have made significant progress in improving synthetic organofluorine chemistry, which can reduce chemical pollution. In another example, visible-light-driven organic reactions greatly lower the energy needed for industrial organic chemical synthesis, wrote Wu Lizhu of the CAS Technical Institute of Physics and Chemistry in the special issue.

The large recruitment of chemists particularly through the Young Thousand Talents Plan has provided a fresh boost to the booming study of chemistry in China, says Zhao Dongbing, a chemist at Tianjin-based Nankai University, who was enrolled as a scholar under the plan after completing his postdoctoral research at Cornell University and the University of Washington in Seattle. China launched the plan in 2008 to attract established scientists and high-tech entrepreneurs to return to China. In 2011, the programme was extended to young scholars, mostly

NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

© 2018 Springer Nature Limited. All rights reserved.



The science behind even basic chemical reactions is hard to prove: a silver chromate precipitation reaction recorded with a macro lens.

YAN LIANG/BEAUTY OF SCIENCE/USTC

targeting highly productive postdocs. By early 2018, China had recruited 3,535 young scientists through the programme.

Zhao believes chemistry accounted for the second largest category of recruits under the plan, after the life sciences, though official numbers are no longer available due to political sensitivities over the recruitment of scholars from the US. Many of the returning postdocs "have good competency, smart ideas and can readily join hot research areas," says Zhao.

These Young Thousand Talent Scholars, including Zhao, Lu and Tan, contribute a large portion of papers from China published in top chemistry journals. For example Xiong Yujie, a chemist at Hefei-based University of Science and Technology of China, who was recruited in the first batch of Young Thousand Talent Scholars in 2011, has since published more than 100 papers, including a dozen in journals tracked by the Nature Index such as *ACS Nano, Advanced Materials,* and *Angewandte Chemie International Edition.*

It is not clear whether the plan will continue to be of as much benefit to chemistry and other disciplines in China, given US accusations that it is a channel for the theft of US technology and intellectual property.

Chinese authorities and universities have removed the name lists of Thousand Talents recruits from websites and some US-based Chinese scientists who might previously have applied have shied away from the scheme.

Although figures for individual research disciplines are not available, China's research and development expenditure has been growing, from 1.03 trillion yuan (US$150 billion) in 2012 to 1.76 trillion yuan in 2017, accounting for 2.1% of its GDP. China's R&D surpasses that of many industrial countries in both absolute and relative terms. For example, in 2016 the R&D/GDP ratio of the United Kingdom was 1.67% when China's was 2.08%.

Surging research investment means that expensive devices, such as infrared and nuclear magnetic resonance spectrometers and elemental analysers, though unaffordable to many chemistry labs in the West, are standard equipment in Chinese research universities.

SUSTAINABILITY WORRIES

Chinese chemistry attracts the same criticism as Chinese science overall: that in the race to publish, China focuses on areas with international traction and quick results.

Zhao acknowledges that the heavy pressure on young principal investigators to publish stops them from delving into the mechanisms underlying basic chemistry, because proving the science behind even very conventional chemical reactions is a long and arduous process.

But, Lu of SUSTech is more optimistic. "An original innovation does not start from scratch, but from following others. Major breakthroughs need gradual accumulations. I have seen more and more Chinese innovative studies that can change our chemistry textbooks." ■

ORGANIC CHEMISTRY

TOP INSTITUTIONS 2015–2017:
1. Shanghai Institute of Organic Chemistry, CAS (Fractional count: 139.7)
2. Lanzhou University (FC: 73.20)
3. Peking University (FC: 67.84)

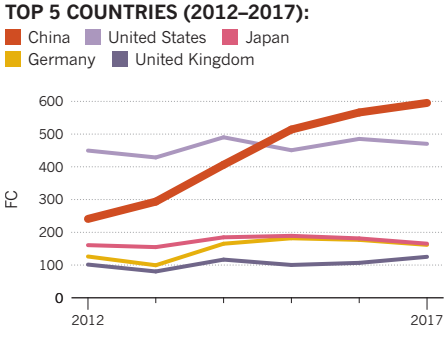
CHINESE INSTITUTIONS IN SUBJECT'S GLOBAL TOP 20: 9

INTERNATIONALLY COLLABORATIVE ARTICLES 2015–2017: 18.7%

TOP CHINESE–INTERNATIONAL COLLABORATIVE PAIR 2015–2017:
University of Science and Technology of China – Nanyang Technological University, Singapore (CS*: 13.96, Article count: 15)

ATTAINED WORLD NUMBER ONE: 2015

TOP 5 COUNTRIES (2012–2017):



*Bilateral collaboration score

NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

© 2018 Springer Nature Limited. All rights reserved.



QUALITY DEFICIT BELIES THE HYPE

COMMENT
FUTAO HUANG

Few Chinese researchers are regarded as global leaders, as the pressure for rapid output prevails.

China's president, Xi Jinping, wants the country to be a world-class innovator by 2050. To achieve this, there are major challenges for the country to overcome — significantly, the quality of its research.

Figures from the United States' National Science Foundation showing China's published science output surpassed that of the US in 2016 have been widely discussed, but other metrics prove celebration is premature.

Data from China's Ministry of Science and Technology suggest that, despite the rapid growth in articles authored by Chinese scholars in the Science Citation Index (SCI) over the decade to 2017, the average number of citations for each article was only 9.4. This is lower than the global average of 11.8, putting China in 15th place by this measure. The SCI tracks articles in high-impact journals.

Tu Youyou, awarded the Nobel Prize in Physiology or Medicine in 2015 for her discovery of a novel therapy against malaria, is the only scientist to have won the prize for research carried out in mainland China. Few Chinese researchers in the 'hard sciences' are regarded as global leaders in their fields compared with researchers in the US, or even Japan, which has much lower output. According to the closely watched Chinese ranking website Netbig, even China's leading laboratories or centres of excellence in such flagship fields as materials research, metals research and chemistry, including those affiliated to the prestigious Chinese Academy of Sciences (CAS), are not ranked among the world's top 10.

Among the reasons is that China's method of evaluating academic research performance values rapid publication output and quick research outcomes over high-quality research with long-term benefits.

My interviews with 19 young researchers and scientists in China over the past two years confirm the pressure they felt to publish articles in SCI journals as quickly as possible. If they don't publish at least half a dozen such articles, and obtain national-level research funding as a principal investigator within their first five years as researchers, they have little hope of being hired as a tenured associate professor or equivalent at a top university, let alone at CAS or the Chinese Academy of Social Sciences.

These evaluation systems have also led to the proliferation of research malpractice, including plagiarism, nepotism, misrepresentation and falsification of records, bribery, conspiracy and collusion. While these problems are not unique to China, the central government's requirement that institutions commit to clear-cut targets for positions in major global ranking systems such as QS or Times Higher Education within a stated time period, mainly by publishing articles in indexed journals, sets China apart. Because institutions and individual researchers stand to benefit greatly from elevating their reputations, no severe punishments have been imposed for academic corruption and malpractice, compared to the US and Japan, although reforms imposing stronger sanctions were announced in May.

Local universities, in particular, view publishing articles in SCI journals as a way





COMMENT
VINCENT LARIVIÈRE, KAILE GONG, AND CASSIDY R. SUGIMOTO

CITATIONS STRENGTH BEGINS AT HOME

The increasing production of papers and a propensity to cite compatriots makes China likely to win the referencing race.

Nations pride themselves on their research activities, using a variety of indicators to size up their scientific capacity and strength against those of close competitors. Spending on research and development, in which China is now second only to the United States, and publication output, in which China overtook the US in 2016, are closely tracked metrics.

These indicators have drawn global attention to the rapid rise of China and its challenge to the science leadership of the US. Our analysis suggests that China is on a trajectory to dominate citations also, due to tendencies to recognize the work of one's fellow citizens.

On citations, the US has long cornered the market, accounting not only for the lion's share but a disproportionate number of the top 1% most highly cited science and engineering papers. While Europe has been slowly gathering a larger share of the highly cited papers, the biggest growth has been from China, whose papers went from below to just above world average in the last decade. If the trajectory continues, China will outpace the European Union in its production of highly cited papers within the next decade or so.

Similarly, there remains a gap in terms of overall citations: the share of all cited references made to US papers, although decreasing steadily since the mid-1990s, remains slightly higher than that of the EU-28, and far above China, which accounts for only about 6% of all cited references.

This gap is likely to close, given the significant role that patriotism plays in referencing behaviour.

SELF-REFERENTIAL

While many words have been devoted to the self-citing practices of individuals, less is known about the degree to which citations stay close to home at the country level. Although the US and United Kingdom garner a large share of global citations, all countries cite themselves disproportionately. This remains true even when controlling for author-level self-citations.

For the EU, as an example, 40% of all references made and citations received are from other European countries. The US, on the

© 2018 Springer Nature Limited. All rights reserved.

to boost their reputation, while gaining bigger budgets, winning research grants and attracting subsidies from local authorities. My interviews with dozens of researchers have confirmed that lesser institutions pay approximately US$10,000 directly to individual researchers for publication of an SCI article. At least one university in Guangzhou pays authors approximately US$70,000 as a reward to go towards their future research for a research article published in *Nature* or *Science*. An author at a top university could receive a bonus of about US$900 for a published SCI article.



CHIAKI ASAHINA ; GUANG NIU / STAFF/GETTY IMAGES ASIA PAC

Chinese scientists face a tough evaluation system.

STRICT CONTROL

The compulsion for Chinese science to serve the needs of economic growth and geostrategic intentions, as well as political ideology, also hinders the creation of an innovative scientific and technological ecosystem. Research supporting sustainable development, including on issues such as climate change, receives comparatively little attention.

Universities and disciplines singled out for special funding under the newly launched Double World-Class project of 2017 are expected to also focus on producing graduates dedicated to constructing a socialist society. There has been increased ideological control over research and more intensive monitoring of internet activity. An expectation that universities adhere to Marxism and take the pronouncements of Xi Jinping as guiding principles, without clear working definitions of the Chinese characteristics institutions should exhibit, restricts scope for innovative thinking, especially in humanities and social sciences.

A 2012 national survey of more than 3,000 full-time faculty members in Chinese universities revealed that only 3% of faculty members aged 31–40 years old are highly satisfied with their jobs. This is the lowest out of any other age group. My more recent national survey with Shen Wenqin of Chinese doctoral students in 2018 indicated that half of them worried about high stress, low salaries, job insecurity and the multiplication of evaluation systems in research. This is especially true in the 'hard sciences'.

According to my separate survey of almost 400 international faculty at Chinese universities in 2016–2018, including interviews with a dozen of them, among the steady increase in scientists coming to work in Chinese universities or research institutes from abroad, very few are top-level scientists, especially not foreign-born and -educated scholars, possibly because they are wary of the working environments described by local academics.

The government's research strategy should be designed to provide young researchers, including doctoral students, with a more favourable academic environment and more promising career future; to reform and improve the present frameworks evaluating scientific research; to let scientists more freely undertake research with international colleagues; and to recognize the role of academic corruption in holding back China's science. ■

Futao Huang *is a professor at the Research Institute for Higher Education, Hiroshima University, Japan.*

other hand, has tended to reference itself much more than the EU: between 1980 and 1996, 60% of references in US papers were to other US papers. However, the US is increasingly citing outside the country and, consequently, receiving a lower share of its overall citations from itself. US self-citations fell steadily from 41% to 33% between 1980 and 2013, though they've since gained a few percentage points to 37%.

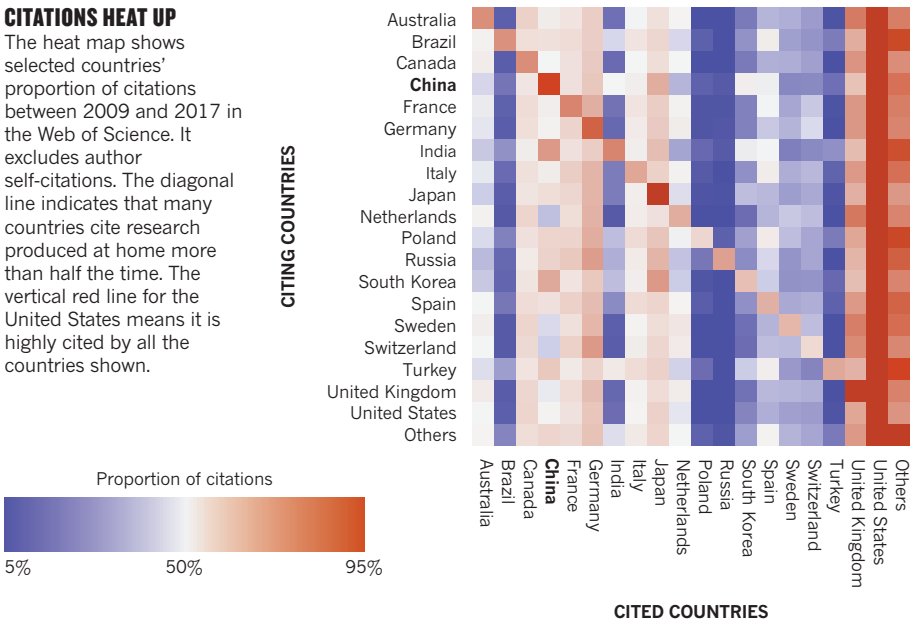
While China's self-references has remained low during the period of assessment, at under 20% of all paper references, Chinese researchers are increasingly citing work from other Chinese researchers. This has resulted in an increase in country self-citations from around 30% in the 1980s to 47% in 2015. China's self-citations as a share of total citations surpassed that of the US in 1999 and is now greater than the US self-citations by 10 percentage points.

The citation payoff is obvious. These trends reveal the tight coupling between paper production and citation: as a country increases production, and therefore a supply of potential references, the country is likely to rely on this capacity and occupy a greater share of the citation space.

Over the past two decades, China has experienced tremendous growth in terms of research investment and production, leading to a larger share of citations. China's increased production is likely to result in a dominance of citations in coming years, and is both a sign and consequence of its rapid maturation as a force in science. ■

CITATIONS HEAT UP

The heat map shows selected countries' proportion of citations between 2009 and 2017 in the Web of Science. It excludes author self-citations. The diagonal line indicates that many countries cite research produced at home more than half the time. The vertical red line for the United States means it is highly cited by all the countries shown.



SOURCE: LARIVIÈRE *ET AL.*

AMÉLIE PHILIBERT; S. CRAIG FINLAY

Vincent Larivière *is an information scientist at the University of Montreal, Canada.*
Kaile Gong *is a PhD student in information science at Nanjing University, China.*
Cassidy Sugimoto *is an information scientist at Indiana University Bloomington, United States.*

© 2018 Springer Nature Limited. All rights reserved.

A GUIDE TO THE NATURE INDEX

A description of the terminology and methodology used in this supplement, and a guide to the functionality available free online at natureindex.com

The Nature Index is a database of author affiliations and institutional relationships. The index tracks contributions to research articles published in 82 high-quality natural science journals, chosen by an independent group of researchers.

The Nature Index provides absolute and fractional counts of article publication at the institutional and national level and, as such, is an indicator of global high-quality research output and collaboration. Data in the Nature Index are updated regularly, with the most recent 12 months made available under a Creative Commons licence at natureindex.com. The database is compiled by Springer Nature.

NATURE INDEX METRICS

The Nature Index provides several metrics to track research output and collaboration. These include article count, fractional count and multilateral and bilateral collaboration scores.

The simplest is the article count (AC). A country/region or an institution is given an AC of 1 for each article that has at least one author from that country/region or institution. This is the case regardless of the number of authors an article has, and it means that the same article can contribute to the AC of multiple countries/regions or institutions.

To glean a country's, a region's or an institution's contribution to an article, and to ensure they are not counted more than once, the Nature Index uses fractional count (FC), which takes into account the share of authorship on each article. The total FC available per article is 1, which is shared among all authors under the assumption that each contributed equally. For instance, an article with 10 authors means that each author receives an FC of 0.1. For authors who are affiliated with more than one institution, the author's FC is then split equally.

The total FC for an institution is calculated by summing the FC for individual affiliated authors. The process is similar for countries/regions, although complicated by the fact that some institutions have overseas labs that will be counted towards host country/region totals.

Two metrics measure collaboration. The multilateral collaboration score (MCS) is an indicator of collaboration between multiple institutions and can be calculated for an individual institution or a group of institutions. MCS takes account of the number of collaborating institutions on a given article, so it can be added across multiple institutions, always resulting in a total FC of 1 for each article.

The bilateral collaboration score (CS) between two institutions A+B is the sum of each of their FCs on the papers to which both have contributed. A bilateral collaboration can be between any two institutions or countries/regions co-authoring at least one article in the journals tracked by the Nature Index.

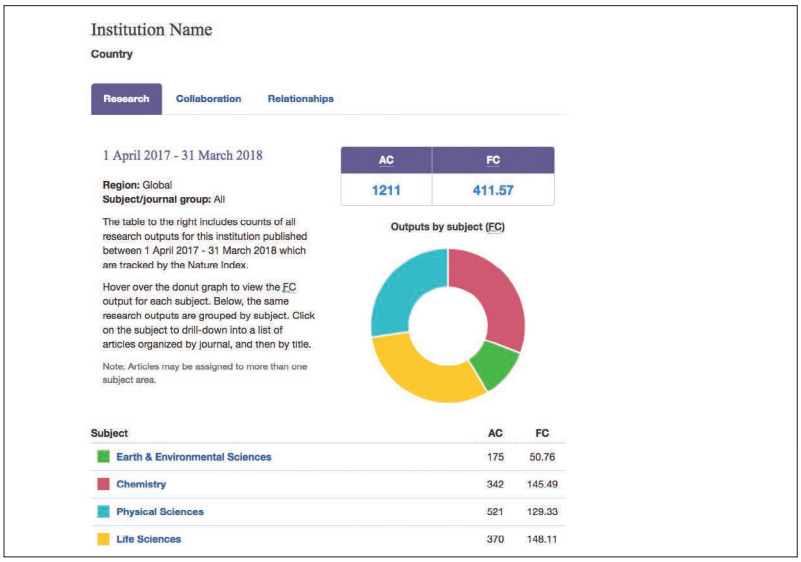
THE SUPPLEMENT

Nature Index 2018 China is based on data from natureindex.com, covering articles published during six years from 1 January 2012 to 31 December 2017 at the country level, and articles from 1 January 2015 to 31 December 2017 at the institution level. Most analyses within the supplement use FC as the primary metric.

The tables rank the top institutions by their FC from 2015 to 2017, overall and according to each of the four broad areas of the natural sciences covered by the Nature Index. Article counts are also included. For the large umbrella organisations, Chinese Academy of Sciences and Chinese Academy of Agricultural Sciences, output has been broken down into their subsidiary institutions. ■

NATUREINDEX.COM

A global indicator of high-quality research



natureindex.com users can search for specific institutions or countries and generate their own reports, ordered by article count (AC) or fractional count (FC).

Each query will return a profile page that lists the country or institution's recent outputs, from which it is possible to drill down for more information. Articles can be displayed by journal, and then by article. Research outputs are organized by subject area. The pages list the institution or country's top collaborators, as well as its relationship with other organizations. Registering allows users to track an institution's performance over time, create their own indexes and export table data.

© 2018 Springer Nature Limited. All rights reserved.

Realizing the promise of gene therapy through collaboration and partnering: Pfizer's view

AUTHORS
Anna P. Tretiakova, John E. Murphy, Michael Binks, Paul Mensah, Joseph Rabinowitz, Douglas M. McCarty, Katherine Beaverson, Molly MacLeod, Gregory LaRosa, and Seng H. Cheng

AFFILIATION
Pfizer, Inc. 235 E 42nd St, New York, NY, USA, 10017

ACKNOWLEDGEMENTS
The authors acknowledge medical writing support from Lazar Partners.

Gene therapy is a promising approach to altering the genetic composition of cells as a way to correct disease-causing mutations or to express proteins or RNA molecules that confer a therapeutic benefit. The concept of gene therapy is straightforward: deliver nucleic acids to target cells to alter their function in a beneficial manner. Moving from concept to reality, however, is a complex process comprised of multiple steps and components, including systems for getting nucleic acids into target cells, DNA regulatory elements that control the amount, location and duration of gene expression, and production of proteins with appropriate activity to alter cellular function in the desired manner. Pfizer is currently focusing on diseases that have single-gene defects, such as certain neuromuscular and hematologic diseases, and we have a robust pipeline of potential gene therapy treatments in preclinical and clinical development. Our current portfolio includes a phase 1b clinical trial for Duchenne muscular dystrophy (DMD), a pivotal phase 3 program in hemophilia B as part of a collaboration with Spark Therapeutics, and an ongoing phase 1/2 trial in hemophilia A in collaboration with Sangamo Therapeutics, Inc.

Given the multiple elements required for successful gene therapy development, collaboration among industry, academia, regulatory, clinician and patient communities is essential. Such partnerships ensure that the necessary expertise is harnessed to achieve maximum benefit: gene therapy products that are clinically beneficial, meaningful to patients and commercially accessible.

Although the clinical utility of *ex vivo* gene therapy has been validated with multiple approved products (for example, Strimvelis, KYMRIAH and YESCARTA), the delivery of genes to cells *in vivo* has been more challenging. One factor that contributes to the challenge is the immune responses that patients may have to viral vectors or to transgene products that were not previously present in the patient's body. Another challenge lies in the limited access that gene therapies may have to the surface of target cells *in vivo* as compared with cultured cells. As a result, gene therapy clinical trials conducted by other pharmaceutical companies over the past 20 years in diverse indications including cystic fibrosis and age-related macular degeneration have failed to advance successfully through clinical drug development.

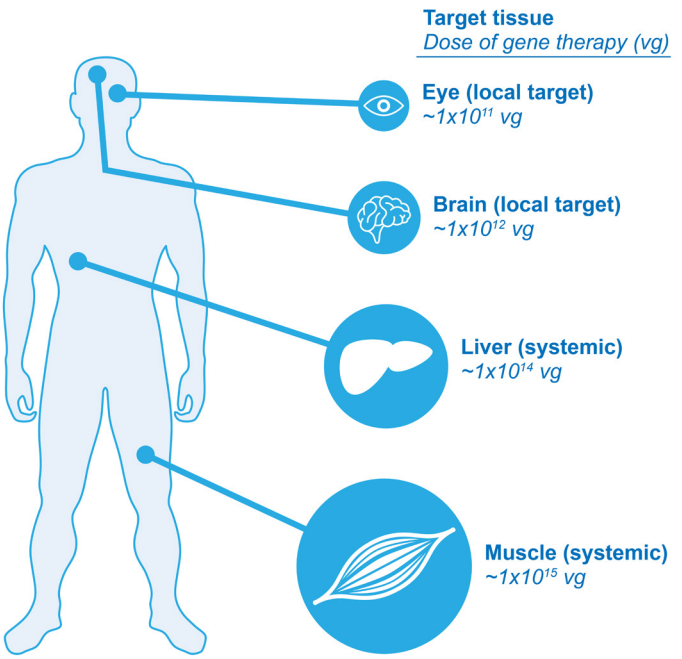


Figure 1: Target size impacts dose. In gene therapy, large target tissues have many cells that need to receive the therapeutic gene, and therefore need to receive a large dose (vector genomes, or vg) of a gene therapy product. For example, gene therapy that is administered as an intravascular injection to provide systemic exposure to multiple organs would need to be given in a large dose. This is in contrast to gene therapy that is administered to a local area of a smaller organ such as the eye, which would require a smaller dose for a smaller number of cells.

However, advances in vector engineering, transgene optimization and the combinatorial use of regulatory elements over the past several years have addressed some of the challenges of *in vivo* gene therapy, and Pfizer believes that gene therapy for single-gene disorders is at a pivotal period in its evolution. In some ways, the field is following a trajectory similar to the development of biologic therapies, which have become common, and critical components of treatment regimens for many serious medical conditions, including cancers, neurologic disorders, diabetes and autoimmune diseases. Innovations in vector design and an enhanced

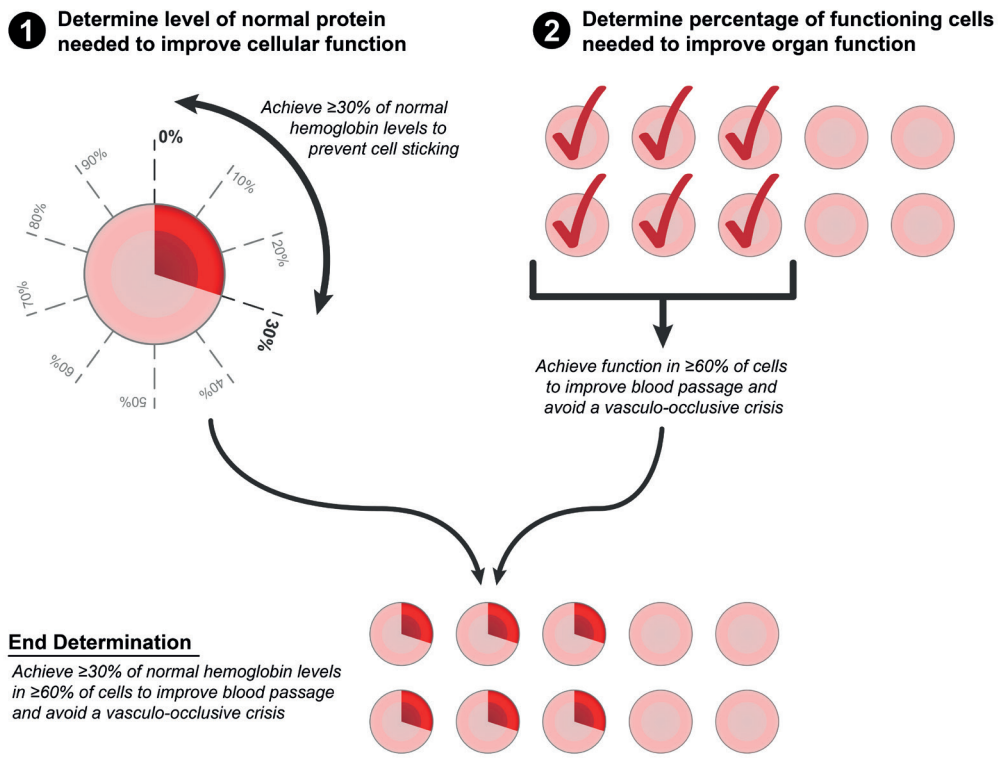


Figure 2: Two components to correcting a disease: Cellular function and tissue or organ function. When considering how to treat a single-gene disease with gene therapy, there are two underlying questions. First: how much protein does the therapeutic gene need to produce in order to improve cellular function? In the example of sickle cell anemia, studies have shown that a red blood cell becomes less sticky if at least 30% of the cell's hemoglobin is not of the sickle variant. Second: how many functional cells does a tissue or organ need to work properly? Transfusion and transplant studies have shown that if at least 60–70% of red blood cells are normal, this is sufficient to improve blood passage and avoid a vasculo-occlusive crisis in patients with sickle cell anemia. Taken together, the answers to these questions provide evidence-based goals when designing a treatment and, in the above example, suggest that sickle cell anemia can theoretically be corrected when at least 60–70% of red blood cells have at least 30% non-sickling hemoglobin[8].

understanding of the biology of certain diseases have enabled the development of novel gene therapy approaches, and numerous late-stage trials of next-generation product candidates, including from Pfizer, are ongoing. Data from these trials will hopefully advance our current understanding of the potential for gene therapy as a therapeutic option, while laying the foundation for future improvements.

The unmet needs of patients with single-gene disorders fall largely in two domains. One area of unmet need is found among patients with diseases for which therapy exists but may only be palliative or associated with high treatment burdens. For example, individuals with hemophilia B are able to address bleeding episodes through frequent intravenous blood transfusions and may be able to achieve disease management with recombinant Factor IX protein therapy. However, gene therapy may improve the quality of life for people with hemophilia B by reducing or eliminating the need for frequent dosing and by ensuring a stable level of Factor IX expression, thus avoiding the peaks and troughs associated with intravenous administration of Factor IX protein. The other type of unmet need is found among individuals with rare, single-gene disorders for which no treatments exist for the majority of patients. Examples in this category include debilitating, progressive disorders, such as DMD, spinal muscular atrophy and Friedreich's ataxia.

The recent technological advances made in gene therapy have opened up the potential to address a broad array of challenges, and Pfizer is pursuing potential solutions in both areas of unmet need. However, achieving the goal of such gene therapies requires not only novel technology, but innovative approaches to solving the technical challenges inherent to the research and development process: translational science, robust and scalable manufacturing, and a collaborative model that takes into account the need to move forward all of these components in tandem. Herein we present Pfizer's perspective on some of the ways in which companies can organize and collaborate in order to work towards realizing the full promise of gene therapy.

TRANSLATING DISEASE BIOLOGY TO THERAPEUTIC STRATEGIES

Focus on understanding disease biology

Although most conditions targeted by gene therapy result from alterations in the activity of a single gene, the effects of these alterations can vary among different cell types and tissues and at various points in a patient's age or developmental stage. Consequently, safe and effective development of gene therapies requires clear insights into disease biology across various cells, tissues and patient demographics. This includes understanding which cell types and tissues need to be targeted and determining if those targets change over time or as a factor of disease progression. While target cell types that undergo frequent cell division require integrating vectors or gene editing strategies for effective treatment, non-dividing target cells can be treated with vectors that remain episomal. Establishing benchmarks for expression of the delivered gene, with respect to both the number of expressing cells and the overall level of expression (**Fig. 1**), is important for informing decisions about optimizing transduction and designing expression cassettes that confer clinically therapeutic levels and localization of the expressed protein. Some diseases, such as DMD, require protein expression predominantly within specific cell types (for example, skeletal and heart muscle). Other diseases, such as hemophilia, may have less stringent requirements for where the protein is expressed as long as it is secreted and at a level that restores correct biologic activity. Access to the tissues to be targeted in a specific disease also plays an important part in vector selection and design, and dosing strategies (**Fig. 2**). For example, in many ongoing clinical studies

that use an adeno-associated virus (AAV) vector to deliver a gene, localized injection of AAV is being used as a dosing strategy to address tissue accessibility.

Hemophilia and DMD, two indications for which Pfizer has potential therapeutics being evaluated in the clinic, provide examples of the role that disease biology plays in gene therapy development. Hemophilia is an indication with a relatively wide therapeutic window with respect to expression levels and relatively straightforward tissue targeting requirements. This is because delivery of Factor VIII (for hemophilia A) or Factor IX (for hemophilia B) coding sequences to liver cells is expected to sufficiently restore blood clotting activity to significantly reduce the bleeding episodes seen in hemophilia. On the other end of the therapeutic range, healthy individuals show levels of Factor VIII or IX up to 150%[1], thus overexpression of the clotting factor is not a concern. The ability to detect Factor VIII or IX activity in the plasma also provides an easily measurable biomarker that is directly related to disease severity and treatment effect. In contrast, DMD represents an indication with additional challenges in effectively transducing a sufficient number of muscle cells and achieving high enough levels of protein expression within those cells to achieve clinically relevant benefit. This is due to the large mass and broad distribution of the muscle tissues that are affected by the disease. The difficulty in accessing muscle cell samples and measuring intracellular dystrophin also complicates obtaining a biomarker of effect.

Freidreich's ataxia exemplifies how disease biology may complicate therapy for a single disease because it affects both the central nervous system and cardiac function. Clinically relevant transduction rates and expression levels may vary among different tissue types and it may be challenging to develop a single gene therapy, or routes of administration, optimized for multiple anatomic compartments.

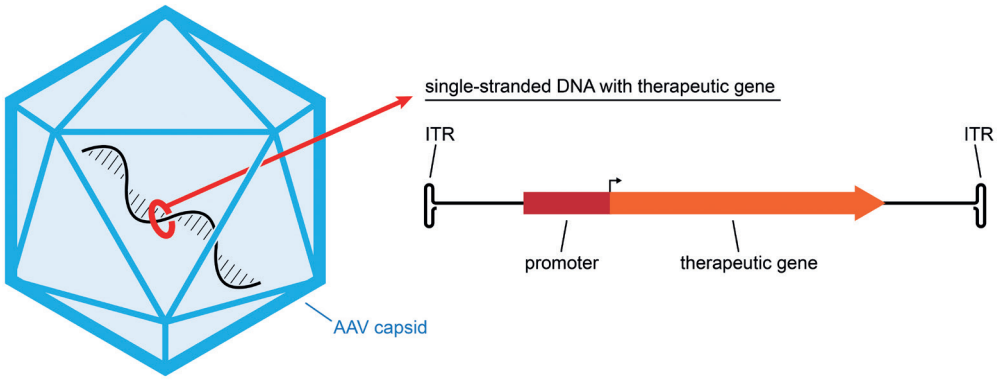


Figure 3: Adeno-Associated Virus (AAV) Gene Therapy Vector. AAV is a small non-enveloped virus that is used to deliver genes to cells in gene therapy. The virus has a linear single strand of DNA inside a protein capsid, and the DNA has two native genes that are used for virus replication. To create a vector for gene therapy, the native genes are replaced with a transcriptional cassette that contains a therapeutic gene and a promoter to direct expression of the therapeutic gene in specific cell types. The cassette also has an inverted terminal repeat (ITR) on each end. After an AAV vector delivers a gene into the nucleus of a cell, and the single-stranded DNA is converted into a double strand, the ITRs are used to convert the linear DNA into a circle, called an episome.

Optimizing vector design and engineering

Given the prominent role that disease biology plays in defining the requirements for potential gene therapies in specific diseases, it is unlikely that a single vector or expression system will be applicable to all indications that could be treated with this therapeutic modality. Thus Pfizer scientists believe it will likely be important to develop a suite of recombinant adeno-associated virus (rAAV)-based vector systems that can be deployed based on the desired product profile. AAV is one of the most actively employed vectors for *in vivo* gene therapies because of its versatility in targeted applications and its safety profile compared to other viruses. Compared to the wild-type AAV, which consists of a capsid shell encasing a single-stranded DNA, the rAAV used in gene therapy lacks components of its own viral DNA but retains the ability to deliver exogenous DNA (that is, an engineered transgene expression cassette) for therapeutic purposes into the nucleus of target cells. Capsid and transgene cassette engineering will therefore play a vital part in developing gene therapies that are optimized for safety and efficacy within specific indications (**Fig. 3**).

Capsid engineering has been used to improve delivery to and transduction of specific tissues, which is necessary for clinically meaningful protein expression. Naturally occurring AAV variants have different tissue tropisms owing to subtle differences in binding preferences of cell surface receptors, and those variants can be of use in targeting specific tissues. Additionally, engineering capsids to develop novel AAV variants, such as through peptide ligand insertion or directed evolution through capsid shuffling, may have further benefits for cell or tissue specificity.

Capsid engineering may also be employed to reduce anti-capsid immune responses that can interfere with transduction activity. AAV is less immunogenic compared with other viruses, primarily because it is unable to replicate or infect without the presence of a helper virus (such as an adenovirus). However, many people have already been exposed to some variant of AAV and may have a pre-existing adaptive immune response to that variant. The concern with such pre-existing immunogenicity is that circulating neutralizing antibodies (NAbs) and T cells may reduce clinical efficacy of AAV vectors. Approaches to overcoming this challenge may include selecting a vector that either has not previously circulated or that does not elicit a clinically significant adaptive immune response, as well as engineering AAV to modify particular antigenic domains on the capsid shell. In any of these cases, performing neutralization assays with living cells to quantify the expected immune response of a given capsid will be helpful in the selection of the vector.

Because a single-stranded transgene delivered to the nucleus of the target cell(s) needs to be converted to a double-stranded molecule to be expressed, this can be a rate limiting step in gene expression[2,3]. One established approach to overcoming this is to develop a self-complementary transgene to bypass the need for single- to double-stranded conversion.

An example of such so-called transgene-cassette engineering is the approach that AveXis (now Novartis) took with the human SMN transgene for the treatment of spinal muscular atrophy (SMA), the leading genetic disease associated with infant mortality (www.avexis.com)[4]. This transgene is under the transcriptional control of the cytomegalovirus enhanced chicken beta-actin hybrid (CB) promoter (NCT03306277)[4]. A pivotal phase 3 trial of this optimized sequence delivered with an AAV9 vector, the same vector Pfizer is using in its DMD clinical trial, is currently ongoing (www.avexis.com; NCT03306277)[4].

The use of naturally occurring gene variants offers another approach to optimizing protein activity[5]. For example, Spark and Pfizer's fidanacogene elaparvovec comprises a naturally occurring variant of Factor IX that has been shown to increase clotting activity eight-fold compared with wild-type Factor IX[5]. Spark has further optimized the codon usage within this Factor IX variant to increase the expressed protein's stability and activity and to reduce its immunogenicity (www.sparktx.com)[6].

Another key component of vector engineering is the development of expression constructs that can fit within the size constraints of particular vector systems. For example, the payload capacity of most AAV vectors is less than 5 kb of sequence, which must comprise coding sequences as well as regulatory elements. DMD results from a loss of functional dystrophin protein. The DMD gene is one of the largest human genes, comprising 2.3 mb of DNA. The full dystrophin cDNA is 14 kb long and cannot be accommodated within an AAV vector. PF-06939926, Pfizer's phase 1 investigational candidate for the treatment of DMD, utilizes a shortened version of the human dystrophin cDNA that has been engineered to provide essential protein function under the control of a human muscle-specific promoter that comprises less than 5 kb of DNA sequence (mini-dystrophin). Similarly, full-length cDNA for Factor VIII is approximately 7 kb of sequence, therefore a fully functional B-domain deleted version that can fit into AAV is being used in all current gene therapy trials.

ROBUST, SCALABLE AND REPRODUCIBLE MANUFACTURING

At Pfizer, we believe that transforming gene therapy from a promising approach to a commercially viable therapeutic modality requires the development of robust, scalable and reproducible manufacturing processes. Consequently, the feasibility of commercial-scale manufacturing of a particular gene therapy candidate must be evaluated at the earliest stages of the development pathway. Manufacturing processes that are being developed by Pfizer also need to balance the goal of establishing a consistent manufacturing approach with the unique vector design requirements for individual disease indications.

One approach that we are exploring at Pfizer is to build internal manufacturing capabilities. This has the potential to provide Pfizer with direct control over process development and flexibility to implement process improvements and to address the parameters of particular gene therapy products.

Another key gene therapy manufacturing asset for Pfizer and other companies pursuing gene therapies is in-house plasmid production capabilities. The availability of high-quality plasmids can often be rate limiting for programs in development. The importance of plasmid quality is highlighted by the potential for the United States Food and Drug Administration (FDA) to place a clinical hold on a gene therapy trial if there is contamination of the plasmid used to manufacture clinical trial material.

Finally, development of robust, scalable manufacturing also requires constant innovation. The desire to innovate proprietary technology, which can be time and cost-intensive, needs to be considered alongside the importance of moving optimized new products toward patients as quickly as possible. Consequently, Pfizer's strategy regarding a given manufacturing technology takes into account its potential clinical value, commercial viability, development time and cost calculations and the company's commitment to improving patient care and outcomes today and in the future.

MULTIPLE APPROACHES TO COLLABORATION

To ensure that translational biology and scalable manufacturing processes advance in tandem to support commercially viable products, Pfizer engages with partners throughout the clinical, research, regulatory, academic and advocacy communities and with smaller, gene therapy-focused biotechnology companies. This is particularly true with respect to rare diseases, for which there are often just a handful of clinicians and researchers with sufficient relevant expertise and hands-on patient experience.

A recent example of the partnership-focused gene therapy ecosystem comes from Spark Therapeutics, Pfizer's partner for hemophilia B gene therapy. Apart from Spark's hemophilia B collaboration with Pfizer, their LUXTURNA gene therapy for confirmed biallelic *RPE65* mutation-associated retinal dystrophy resulted from 25 years of research by Jean Bennett, PhD, and Al Maguire, MD, at Penn Medicine's Center for Advanced Retinal and Ocular Therapeutics group in collaboration with the Children's Hospital of Philadelphia (CHOP) in the United States (www.pennmedicine.org). The 2013 spinoff of Spark from CHOP enabled the company to develop its landmark gene therapy trial that led to its 2017 FDA approval (sparktx.com; www.blindness.org)[7]. Similarly, research from the network of academics and clinicians at the Nationwide Children's Hospital's Center for Gene Therapy in Columbus, Ohio in the US has led to the spinoff of several companies with gene therapies in clinical trials, including AveXis.

Seamless integration of technological expertise and resources across key communities should enable an informed development process that can reduce the time and cost needed to develop therapies that truly address patients' clinical needs and quality of life concerns.

Consortia participation

Another way in which Pfizer collaborates with other key gene therapy communities is through participation in a variety of consortia. Pfizer has participated in many consortia, including public–private partnerships such the Innovative Medicines Initiative (IMI), a European Union initiative that seeks to address specific, yet widespread healthcare challenges by facilitating the sharing of knowledge and resources among cross-disciplinary partners. We believe there is a significant opportunity to leverage this collaborative model for gene therapy, such as through IMI's

'Think Big' initiative for advanced therapy medicinal products (ATMPs). This collaborative approach encourages the development of solutions to challenges associated with this therapeutic paradigm.

Engagement with patient advocacy organizations and patient communities

Individuals with rare diseases and the advocacy organizations that represent them have expertise in the 'lived experience' of their disease. They are often experts in the science and biology of their diseases and are focused on building their capabilities to engage with the medical and pharmaceutical industry communities in the drug development process. They can provide invaluable insights about the burden of disease and the impact of current therapies on their lives. These insights are critical in establishing a true patient-centric approach that values benefit, meaningfulness and access. This is particularly true with emerging technologies such as gene therapy, where uncertainty is a variable that must also be factored into any risk–benefit equation and patient choice. Patient advocacy associations can be quite impactful in educating regulators about risk–benefit profiles that the patient community views as acceptable, motivating patients and providers to participate in clinical trials and enabling scientific advancement through collaborations with academia and industry.

Opportunities for collaboration with patient and advocacy groups may differ from one community to another based on the unmet need and priorities. For patient communities in which gene therapy is advancing through clinical development and for whom there is historical clinical research experience (for example, DMD), opportunities for collaboration with industry may focus on community education, community engagement to optimize development programs, and/or creating shared expectations on what gene therapy can and cannot achieve for disease management. Pfizer works closely with several advocacy groups within the DMD community on coalition-model efforts to solve for common challenges in drug development, including the Collaborative Trajectory Analysis Project (cTAP: http://ctap-duchenne.org), the Duchenne Regulatory Science Consortia (D-RSC: https://c-path.org/programs/d-rsc) and Project HERCULES (HEalth Research Collaboration United in Leading Evidence Synthesis: http://hercules.duchenneuk.org). Pfizer is also advancing with Parent Project Muscular Dystrophy (www.parentprojectmd.org) important patient preference research seeking to obtain quantitative evidence on patient and caregiver views on benefit and risk of emerging therapies including gene therapy.

For patient communities where gene therapy may be in earlier pre-clinical or even discovery phases, opportunities for collaboration with industry may focus more on transparency about the research agenda, optimizing access to scientific thought-leadership and data, and advancing the science through de-risking strategies, such as cost-sharing and consortia.

Harmonizing strategies for drug approval and market access

Pfizer has been a part of ongoing conversations with regulators and payors that we believe are essential for harmonizing approaches for drug approvals and market access. A clear regulatory pathway and a regulatory environment that values gene therapy as a potentially life-changing treatment modality are critical for the success of the field as a whole. Regulatory agencies, especially the FDA under commissioner Scott Gottlieb, have recognized these needs and have recently drafted guidance for pre-clinical, clinical and chemistry, manufacturing and controls expectations.

Following the issuance of the draft guidance, Gottlieb made clear the need for both consistency and flexibility in harmonizing gene therapy regulatory guidelines (www.fda.gov). Consistent and clear regulatory expectations will be critical for enabling gene therapy as a robust therapeutic class with commercially viable timelines and development costs rather than as a collection of one-off products that each require a new, expensive and time-consuming regulatory process. This is of particular importance for the development of gene therapies for small patient populations to ensure economic viability. And as regulatory agencies seek to exercise flexibility to facilitate gene therapy development, there also remains a lack of consistency in where to offer such flexibility, which creates a challenge for global development. Companies such as Pfizer that are pursuing a global plan for product development will be well positioned to help drive common experience and shared learning among regulatory agencies.

PFIZER'S AIMS FOR FUTURE INNOVATION

After decades of development, Pfizer believes that gene therapy for single gene disorders is on the cusp of becoming a robust therapeutic modality in a variety of disease indications, and it is a major focus of our efforts in rare disease. Additional advances in vector engineering, disease modelling and both clinical and commercial-scale manufacturing are essential to ensure that the full potential of this approach is realized for as many patients as possible. We hope that our continued innovation and collaborations with academia, industry and patients will allow us, and others in the field, to transform gene therapy from an interesting scientific concept into a broad portfolio of commercial products that improve patient care and outcomes.

REFERENCES

1. World Federation of Hemophilia Report on the Annual Global Survey 2015 (2016). Available from: http://elearning.wfh.org/resource/annual-global-survey-2015 [Accessed Oct 10 2018].
2. Ferrari, F. K., Samulski, T., Shenk, T. & Samulski, R. J. Second-strand synthesis is a rate-limiting step for efficient transduction by recombinant adeno-associated virus vectors. *J. Virol.* **70**, 3227–3234 (1996).
3. Fisher, K. J. *et al*. Transduction with recombinant adeno-associated virus for gene therapy is limited by leading-strand synthesis. *J Virol*. **70**, 520–532 (1996).
4. Mendell, J. R. *et al*. Single-dose gene-replacement therapy for spinal muscular atrophy. *N. Engl. J. Med*. **377**, 1713–1722 (2017).
5. Arruda, A. & Samelson-Jones, B. J. Factor IX Padua: from biochemistry to gene therapy. *Blood* **128**, SCI-9 (2016).
6. George, L. A. et al. Spk-9001: adeno-associated virus mediated gene transfer for hemophilia B achieves sustained mean factor IX activity levels of > 30% without immunosuppression. *Blood* **128**, 3 (2016).
7. Jameson, J. L. Communication from the Dean: FDA approval of gene therapy for blindness. Perelman School of Medicine, Univ. Pennsylvania (2017). Available from: https://www.med.upenn.edu/evpdeancommunications/2017-12-19.html [Accessed Oct 10 2018].
8. Steinberg, M. H., Chui, D. H. K., Dover, G. J., Sebastiani, P. & Alsultan, A. Fetal hemoglobin in sickle cell anemia: a glass half full? *Blood* **123**, 481–485 (2013).

INDIA HAS DESIGNS ON TOMORROW'S SCIENCE ECOSYSTEM TODAY

A conversation with **KRISHNARAO APPASANI**, vice-chair and chief executive officer, Science City of Andhra Pradesh



Science and technology are vital to India's development and rapidly growing economy. Considering its size, India doesn't count many scientists in its population, and home-grown researchers often take positions elsewhere. To move from this 'brain drain' to 'brain regain', Harvard and Massachusetts Institute of Technology (MIT)-trained molecular biologist Krishnarao Appasani has returned (after three decades in the USA) to his home state to make real his vision of having three science cities in Andhra Pradesh, each aiming to attract young Indian people into science and provide opportunities for them to stay.

What is the inspiration for these science cities?
The idea started in December 2014 when I met with the then Union Minister of State for Science & Technology & Earth Sciences, Y. S. Chowdary. There are science cities elsewhere in India. However, while these encourage and promote science, they do not include a platform for innovation-based education and research.

After the minister and I discussed science cities 'from concept to execution' over many meetings and presented to his secretaries, they recommended me to the chief minister of Andhra Pradesh, Nara Chandrababu Naidu. In June 2016, Naidu recruited me to head up the project under his stewardship. It will be a non-profit, government statutory autonomous body supported by the state and the central government. Alongside their funding, we will also raise money, and seek sponsorship from philanthropists, once we have a detailed masterplan.

How are the plans taking shape?
My vision is to develop the first science city infrastructure in India, creating a hub for science and technology in buildings designed by the world's greatest architects. I am excited that the 'science-centric and eco-centric' masterplan will be developed by an architect who is a son and brother of Nobel laureates. I believe that Andhra Pradesh could be the leading region in India for science by 2025, which could boost India's reputation worldwide. If successful, these science cities could extend across India.

The first science city will be in Tirupati, with a further two in Amaravati and Visakhapatnam. The structure will be similar across each: a science-promotion cluster with seven museums; an innovation cluster including research institutes and incubation centres in biomedical and physical sciences; and a science-dissemination cluster, with a convention centre and accommodation including around 300 bedrooms. The convention centre will initially house around 5,000 people, with potential to expand to 20,000. Its main role will be for scientific congresses, but it could also be used for large social gatherings and other celebrations, to generate revenue. All buildings will be created to high environmental and sustainability standards. I haven't seen this model anywhere else in the world.

I BELIEVE THAT ANDHRA PRADESH COULD BE THE LEADING REGION IN INDIA FOR SCIENCE.

Why do seven museums feature in the designs?
Tirupati is home to the Tirumala Venkateswara Temple, one of the holiest Hindu pilgrimage sites, and has around 100,000 visitors every day. It is on the Seshachalam Hills, whose seven peaks are said to represent the seven heads of Adisesha. In honour of this, we are creating seven different museums around Tirupati, each designed by a different world-class architect representing a different country. Each museum will have a different focus: arts and sciences; air, space and defence; media; transport; human evolution and anthropology; rainforests and biodiversity; and a children's museum and planetarium developed with an astrophysicist.

How are you involving the next generation?
In 2014, part of Andhra Pradesh was separated off to create Telangana state, which includes the capital Hyderabad, losing us a lot of scientific talent. Because of this, and the emigration of many scientists to the USA and Europe, we want to encourage and educate a new generation of young scientists. The many pilgrims to the temple in Tirupati include children and young people, and we hope that they will visit our museums, learn about science, and perhaps think about choosing science as a career. This includes supporting the education of girls and young women in science; we want to improve the gender ratio.

What is needed to make these dreams become reality?
Now that the concept is approved by government, the next step is to create the masterplan for the first science city, which we expect to complete within the next six months, and then we can start building. We aim to have this finished in the next five years.

I would hope to have all three science cities up and running ten years from now. The timing for Amaravati and Visakhapatnam may depend on the political arena, as there will be assembly elections in Andhra Pradesh in April or May 2019, but I am confident that whichever party gets in, this project will continue to create the first 'science ecosystem' in the nation.

Science City of Andhra Pradesh

ADVERTISER RETAINS SOLE RESPONSIBILITY FOR CONTENT

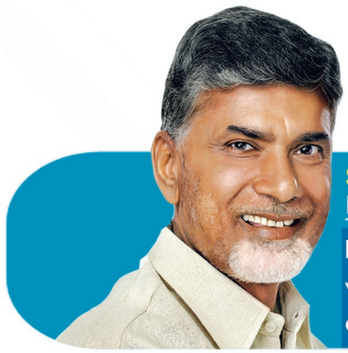
Science City of Andhra Pradesh

SCIENCE CITY OF ANDHRA PRADESH

'BOLD IDEAS TO INSPIRE, INNOVATE AND LEAD THE FUTURE'

TIRUPATI | AMARAVATI | VISAKHAPATNAM

"We aim to build the Role Model Science City in the country which will become a hub for future Science Discoveries and Technology Innovation in India".





SHRI. NARA CHANDRABABU NAIDU
Hon'ble Chief Minister of Andhra Pradesh & Chairman, Science City

Development of 'Science City Infrastructure' is a part of CM's vision to make Andhra Pradesh a knowledge hub for education and research.

Leading a New Era In Science Promotion & Research



THE SCIENCE CITY OF ANDHRA PRADESH (AT TIRUPATI) CONSISTS OF THREE PROGRAMS/CLUSTERS

1. Science Popularization Cluster
   Science Museums.
2. Knowledge Dissemination Cluster
   Convention Centre.
3. Science Discovery Cluster
   R&D Innovation Centre.

We planned to constitute multi-ministry sponsored institutes like:

- International & Inter-University Institute for Quantum Science and Technology (IIQST) to be sponsored by Department of Science & Technology (DST), MoST, GoI.
- National Institute of Metabolomics & Diagnostics (IIMD) to be sponsored by Department of Biotechnology, MoST, GoI.
- Indian Institute of Robotics, Automation & Artificial Intelligence to be sponsored by Council for Science & Technology Industrial Research (CSIR), MoST, GoI.

To know more about the SCIENCE CITY OF ANDHRA PRADESH
VISIT: **www.sciencecity.ap.gov.in**

nature

[insideview]

MANIPAL
ACADEMY of HIGHER EDUCATION
(Deemed to be University under Section 3 of the UGC Act, 1956)

Profile Feature as seen in *Nature* 13th December 2018

AN EMINENT INSTITUTION REACHING FOR THE SKY

A conversation with **VINOD BHAT**, vice chancellor, Manipal Academy of Higher Education



The Manipal Academy of Higher Education (MAHE) is a rapidly expanding institution with a bright future ahead. The university, whose medical school was the first private-sector educational institution to be opened in India back in 1953, has recently been awarded the title of 'Institution of Eminence' by the Indian government. The university has five campuses: three in India, one in Dubai and one in Malaysia. The government accolade will help MAHE to soar as it maintains its course in becoming a world-leading institution.

Where does MAHE have its sights set?
Our focus is on global leadership in human development, and the delivery of world-class education in all subjects. Our goals very much stem from our history as a leading private-sector education specialist in India. Our medical school, founded by Dr Tonse Pai, proved so successful that other schools soon followed, and we diverged into Health Sciences, Dental Sciences and Engineering. We gained university status in 1993. The quality of teaching at MAHE has been a continuous source of inspiration for our students and staff, and our ethos promotes teamwork, inclusivity and integrity. A review of progress in 2005 led the university into a transitional phase, progressing from a teaching university towards strengthening its research capabilities. We haven't looked back since.

In which research areas is the institute flying high?
We remain a key institution for medical science and dentistry. Since 2005, we have expanded our research in health sciences to include life sciences, molecular biology and drug discovery. We have exchange programmes with our twinned universities in the USA, UK, Australia, and other countries, allowing our students and researchers access to the international academic community.

We have a thriving public-health department who work very closely with communities and policy-makers to improve responses and treatments for the health challenges in our country. Indeed, earlier this year researchers from our Department for Virus Research guided the response to the Nipah virus outbreak in Kerala at the start of the rainy season. Thanks to the MAHE team working closely with local authorities and health professionals, the epidemic was identified early, and the outbreak was quickly contained. We also have several basic and translational research programmes focusing on human diseases.

Another of our main research areas is engineering, and our avionics research team are currently working alongside the Indian government on various projects. Our Centre for Humanities is also flourishing.

How has MAHE risen above the crowd?
Through our past performance as a high-quality teaching university, we have shown ourselves to be high achievers. Several of our schools are in the top-ten institutions for their subject in India. Our recent success in transitioning from purely teaching to a fully-fledged research-orientated university has confirmed our commitment. Our research teams are on course to produce 3,000 publications a year, up from 500 just a few years ago.

SEVERAL OF OUR SCHOOLS ARE IN THE TOP-TEN INSTITUTIONS FOR THEIR SUBJECT IN INDIA.

Will this award act as a launchpad for new initiatives?
The higher education sector is tightly regulated in India, and the first benefit we will see immediately is that we are freed from these regulators. This means we can experiment with the latest teaching and learning tools, revise and implement new pedagogy as we see fit, and hire staff from overseas without any red tape in the way. This is a significant moment and provides us with so many exciting opportunities for the future.

The eminence award has already allowed us to set new goals for the coming years. We aim to scale-up our research output and publications and expand our faculties and postgraduate student numbers. By 2022, we aim to enrol 1,000 new PhD students every year. We aim to be one of the top 500 universities in the world within ten years.

What makes MAHE a unique space for study and work?
The small university town of Manipal, where we have our main campus, is an idyllic place. It is set in beautiful countryside in Karnataka's Udupi district, and we are lucky that many of the townsfolk are either involved, or have been involved, in the university — many as students, many as faculty staff. This gives MAHE a close-knit, community feel. Everyone lives and works within the campus. In Manipal, we commute to work on foot and breathe the fresh air of the countryside rather than traveling across a busy city each day.

We have an excellent school for our children and a first-class hospital on site. The families of staff are well catered for.

What would you like to say to prospective students and researchers seeking new horizons?
Do come for a visit and explore; make this lovely campus part of your studies! We have excellent facilities for young researchers at all stages of their career, a good mentoring programme with experienced researchers, and a multitude of curricular and extracurricular activities ongoing. Come to MAHE, and you too can be part of our bright future.

MANIPAL ACADEMY of HIGHER EDUCATION
(Deemed to be University under Section 3 of the UGC Act, 1956)

ADVERTISER RETAINS SOLE RESPONSIBILITY FOR CONTENT



Manipal Academy of Higher Education (MAHE) is celebrating becoming an Indian Institute of Eminence (IOE). MAHE started as a medical education establishment in 1953, and has remained at the forefront of its field ever since. The government of India converted MAHE into a university in 1993 — the first medical sciences institution in the private sector to be accorded this recognition — and the IOE designation comes in its silver jubilee year. MAHE offers bachelors, masters and doctoral degrees, and has institutes that excel in such areas as medicine, dentistry, engineering, life sciences, nursing, allied health, pharmacy, management, communication, information sciences and hotel management. The university has state of the art research facilities in the area of basic, applied and translational research, and is focused on trans-, inter- and intra-disciplinary research.



MAHE's consistent emphasis on quality has seen its degrees being recognized worldwide, ensuring collaboration with reputed national and international universities. The university cultures a strong educational and research environment, and can draw on cumulative experience gained over six and a half decades of producing thousands of graduates, an excellent academic reputation, experienced faculty, and excellent academic, research and clinical facilities.



Research Positions Open

MAHE invites applications for a) faculty, b) post-doctoral fellows and c) researchers leading to PhD degree. MAHE particularly encourages applications from faculty and post-doctoral fellows who have excelled in research in biological sciences related to human health/biomedical research and who have a proven track record to drive independent research programmes. MAHE is also hiring individuals who have expertise in biosensors, biomedical devices, bioprinting, bioengineering, microfluidics, tissue engineering, stem cell research, neurosciences, systems biology, infectious diseases, immunology, genomics/genetics/epigenetics and pharmaceutical sciences. Positions and remunerations are commensurate with experience.



For further information, please contact:
Deputy Director-HR
Manipal Academy of higher education, Manipal, Manipal-576104
email: jobs@manipal.edu; hr.sols@manipal.edu.

Springer Nature

Make your mark in the city of makers

Calling talented faculty to the
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN (HITSZ)



The 'City of Creators', Shenzhen is China's frontrunner in promoting innovation-driven development. Harbin Institute of Technology, Shenzhen (HITSZ) maintains high standards to attract excellent, international talents, while carrying forward the innovation spirit of Shenzhen to contribute to national and regional economic and social development.

Harbin Institute of Technology (HIT), founded in 1920, is a national key university under the Ministry of Industry and Information Technology. It offers specializations in science, engineering, management and many other fields. It is a member of the C9 League and one of the first universities to be selected for the national Project 985. It became part of the national Double First-Class initiative in 2017.

HIT was ranked as the world's sixth best university for engineering, and the second in China, according to the 2018 US NEWS Global Universities ranking. In the Academic Ranking of World Universities (ARWU) 2017 ranking, it was eighth in China and among the global top 200. In the same year, HIT had a subject area ranked among the global top 0.01%, according to Essential Science Indicators (ESI) data.

Together with the Shenzhen Municipal Government, HIT created Harbin Institute of Technology, Shenzhen (HITSZ) in 2002. It is now one of its key campuses. It is the first university among the C9 League to open a campus and enrol undergraduates in Shenzhen.

For more details about HITSZ, please refer to www.hitsz.edu.cn.

HITSZ now has multiple faculty positions available. It is eagerly seeking talented researchers from around the world to join its dynamic team in Shenzhen. ■

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

FIELDS OPEN FOR RECRUITMENT:

1. Computer Science and Technology
2. Electronic Science and Technology
3. Materials Science and Engineering
4. Control Science and Engineering
5. Power Engineering and Engineering Thermophysics
6. Mechanics
7. Mathematics
8. Practical Economics
9. Business Administration
10. Civil Engineering
11. Mechanical Engineering
12. Architecture
13. Space Science and Technology
14. Information and Communication Engineering
15. Environmental Science and Engineering
16. Management Science and Engineering
17. Biomedical Engineering
18. Chemistry
19. Physics
20. Biology
21. Design
22. Sociology
23. Marxism
24. Marine Science
25. Aeronautical and Astronautical Science and Technology
26. Urban and Rural Planning
27. Linguistics
28. English

QUALIFICATIONS AND REQUIREMENTS:

1. A Ph.D. in a related field;
2. Experience working abroad or postdoctoral training is preferred

SALARY AND BENEFITS:

1. Successful applicants will be appointed as 'Professor', 'Associate Professor' or 'Assistant Professor' according to their qualifications and backgrounds;
2. An annual salary in the range of 300,000 to 1.5 million RMB;
3. Research funds will depend on position and field.
4. Support to apply for high-level talents allowance in Shenzhen (ranging from 1.6 to 3 million RMB);
5. Subsidized housing

TO APPLY:

Applications must include the following documents:

1. Application Form for Faculty Position (downloadable from: http://www.hitsz.edu.cn/job/view/2.html; please indicate main research areas to facilitate the application process);
2. A cover letter in three parts: an introduction explaining why you should be considered for the job, expected contributions to the school in terms of research, and future work plan;
3. Three letters of recommendation
4. Electronic copies of supporting documents, including diploma, lists of publications and achievements, etc.

Application materials should be sent to YANG Zhixi at: hrsz@hit.edu.cn.

CONTACT:

Ms. YANG Zhixi

Human Resources Department
Harbin Institute of Technology, Shenzhen
E-mail: hrsz@hit.edu.cn.
Tel: +86-755-26033365

A NON-STOP ROUTE TO COLLABORATIVE DISCOVERY

In line with the Belt and Road Initiative, SHANGHAI IS BROADENING ITS INTERNATIONAL REACH and seeing its research and technologies taken up across the world.

In Minsk, the capital city of Belarus, a 12km bus route is drawing attention. Developed by China's Shanghai Aowei Technology Development Company, it is designed for supercapacitor-powered electric buses, which use 14% less energy than conventional electric powered buses, while reducing the cost by 6%. The bus line, operational since early 2017, is the first project to emerge from the China-Belarus Industrial Park, a special economic zone 25km east of Minsk. Set up under an intergovernmental agreement, the industrial park, called Great Stone, is an offshoot of the Silk Road Economic Belt.

The Belt and Road Initiative, launched by the Chinese government in 2013, aims to improve regional cooperation and connectivity by strengthening infrastructure development, trade and investment in countries along the ancient Silk Road and beyond. In support of this national initiative, the municipal government of Shanghai is launching diverse programmes to broaden international exchange and promote collaborative innovation, pushing the depth, range and quality of international collaboration to the next level.

Supporting technology flows from China

Aowei is one of the first companies to be registered in the China-Belarus Industrial Park. Aowei developed the world's first supercapacitor bus and its commercially operated bus route. Its supercapacitor technology, quickly rechargeable and energy efficient, offers an inexpensive and environmental friendly solution to city transportation. "We are proud to bring our newest technologies out of the country to benefit people across the world," said Hua Li, Aowei's chairman. "This is also in line with the objectives of the Belt and Road Initiative, as technology transfer stretches outside national borders."

Within the framework of the national initiative, the company is supported by the Science and Technology Commission of Shanghai Municipality (STCSM) to open bus lines in Serbia, Bulgaria and Austria, in addition to Belarus. "With a production and R&D base in the China-Belarus industrial park," said Hua, "we have a gateway to enter the European market."

In Israel, Aowei has collaborated with local companies to build integrated electric bus power systems, design routes, and support power supply infrastructure construction. Joint efforts have led to a supercapacitor model bus line with charging stations, and technology standards for supercapacitor buses, which were issued in Israel as national standards. It was a first for China to see its technology standards in new energy vehicles adopted in a developed country. "We appreciate the government support facilitating this," said Hua.

As Shanghai is speeding up its development into a global science and technology innovation centre, local enterprises have gained tremendous opportunities for development and enhancing their innovation capabilities, said Zhang Quan, director general of STCSM. "We have set the scene for local enterprises to reach out to the world and for promoting technology transfers at home and abroad." The municipal government has broadened Shanghai's international collaborations, providing support to local enterprises for transferring technology and establishing R&D centres overseas. "We encourage our enterprises to collaborate with foreign partners on industrial park construction, as well as R&D centres and other projects," said Zhang. "We look to transform technological innovations into drivers of productivity, and enterprises are the main actors in the process."

Lotusland Renewable Energy Technology (Shanghai) is another high-tech enterprise in Shanghai to reach out via the Belt and Road Initiative. Specializing in geothermal energy, the company, also registered in the China-Belarus Industrial Park, is working with the Belarus government to develop geothermal energy and other new energy, as well as infrastructure building and manufacturing. Several projects have set the scene for collaboration in the industrial park, contributing to exploring new models of collaborative innovation, particularly in the field of energy.

Deepening research collaboration

The Shanghai government's efforts to promote science collaboration with Belt and Road Initiative countries deepen cultural exchange, particularly among young researchers, and support construction of joint technology platforms to enhance research collaborations.

Following a Ministry of Science and Technology plan, STCSM initiated an exchange programme in 2017 that funds bright young scientists from countries along the Silk Road Economic Belt to spend six to 12 months in Shanghai for study and work. From 2017 to 2018, with a budget of 27 million RMB, the programme has supported 90 young scientists from Belt and Road Initiative countries to come to Shanghai. Financial support from STCSM will continue for the next few years.

"Talent is the basis and link for regional cooperation on science and technology innovation," said Zhang. With the growing research capacity of Shanghai, the city's attraction is growing. "Globally, more bright young researchers are setting their sights on the city."







Delegates of the Belt and Road Initiative science and technology innovation consortium gathered in Shanghai for its first summit.

To deepen research collaboration, STCSM has also launched a joint programme to build around 20 laboratories or research centres within five years with countries included in the Belt and Road Initiative. "By consolidating and sharing resources, we hope to boost both parties' research capacities and address common challenges," said Zhang.

The first phase of the project has supported joint laboratories on neurological diseases, precision chemistry, and Terahertz technology. These laboratories will join forces to improve research, talent training and technology transfer, promoting science and technology exchange between Shanghai and partnering countries.

For example, based on the established collaboration between the governments of Shanghai and the New Zealand city of Dunedin, Shanghai's Huashan hospital has joined forces with the New Zealand-China Non-Communicable Diseases Collaborative Research Centre to build a joint laboratory on neurological diseases.

The project has received 1.5 million RMB from STCSM and has led to the establishment of a comprehensive brain sample bank, the first of its kind in Shanghai.

In search of new cooperation mechanisms, STCSM has also organized international forums that gather government officials, thinktank participants, and leaders of industry.

Led by Shanghai Jiaotong University, a Belt and Road Initiative science and technology innovation consortium has been established, providing a resource-sharing platform to maximise the benefits of collaboration. ■